Learning to Recognize Faces from Videos and Weakly Related Information Cues *

Martin Köstinger, Paul Wohlhart, Peter M. Roth, Horst Bischof
Institute for Computer Graphics and Vision, Graz University of Technology
{koestinger,wohlhart,pmroth,bischof}@icg.tugraz.at

Abstract

Videos are often associated with additional information that could be valuable for interpretation of their content. This especially applies for the recognition of faces within video streams, where often cues such as transcripts and subtitles are available. However, this data is not completely reliable and might be ambiguously labeled. To overcome these limitations, we take advantage of semi-supervised (SSL) and multiple instance learning (MIL) and propose a new semi-supervised multiple instance learning (SSMIL) algorithm. Thus, during training we can weaken the prerequisite of knowing the label for each instance and can integrate unlabeled data, given only probabilistic information in form of priors. The benefits of the approach are demonstrated for face recognition in videos on a publicly available benchmark dataset. In fact, we show exploring new information sources can considerably improve the classification results.

1. Introduction

The vast amount of digital video data that is constantly made available by TV and video-sharing websites could be an extremely valuable and important source of information. However, this data is hard to access, since it is mainly indexed by some meta-data and not by its content. Automatic methods interpreting the visual content would be beneficial to allow for a more efficient search. In this work, we are particularly interested in fully automated identification of people in videos, requiring to solve the following challenges. First, detecting people (*i.e.*, their faces) and tracking them throughout a scene. Second, describing their appearance for a later re-identification. Third, extracting information from associated cues such as the audio track (speech recognition), subtitles, the transcript, on-screen text, or electronic program guide (EPG) data.

This problem was recently tackled by several authors [1, 2, 5, 6, 11, 14]. Everingham *et al*. [5, 6] label exemplars by visual speaker detection. The name of 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*. [14] replaces the nearest neighbor framework by multiple kernel classification. The base kernels operate on the min-min distance between HOG [3] blocks. Therefore, the optimized combination coefficients describe the relative importance of the individual blocks for classification. Nevertheless, the hard-labeling cannot deal with unreliable information. Moreover, it is not possible to integrate cues providing information that cannot be assigned unambiguously to one single instance. Ramanan *et al*. [11] use a multitude of inference cues to obtain face clusters. The cues apply to different time scales. However, the system requires manual user interaction to label an initial set of face clusters.

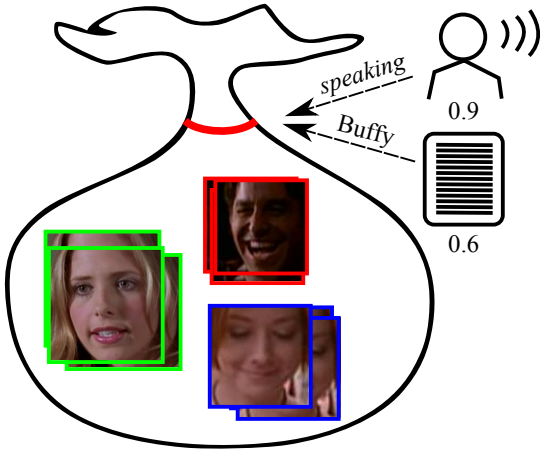


Figure 1: Face recognition in videos as SSMIL problem: Multiple instance learning enables to incorporate ambiguous information (*e.g.*, the video transcript reveals that a character is present, but the corresponding face is unknown). Semi-supervised learning makes use of not fully reliable information. (*e.g.*, a character is speaking with a certain probability).

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

978-1-4577-0845-9/11/$26.00 ©2011 IEEE

Thus, these methods require either manual labeling, cannot integrate unreliable information, and information that applies to multiple instances cannot be used. However, these are reasonable scenarios when learning from videos and associated sources, as illustrated in Figure 1. For instance, we know from textual cues that a specific character should be present in a video scene. But we do not know to which face it corresponds or even if it is visible.

The goal of this paper is to inherently deal with noisy and uncertain labels and use of information which cannot be disambiguated. In particular, we meet these requirements by proposing a new Semi-Supervised Multiple Instance Learning (SSMIL) algorithm. Semi-Supervised Learning (SSL) allows to incorporate labeled and unlabeled data. A special case is to include probabilistic prior information about the unlabeled samples. Another significant problem is that for supervised learning each sample needs to be given a label. This is often either hard or even impossible. 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) [4]. In MIL data is provided in form of labeled bags, where a bag is positive if *at least one* instance in the bag is positive whereas accordingly for a negative bag all instances are negative.

Hence, it is clear that both approaches could be beneficial for the given task. On the one hand the associated information cues provide priors which can be used in an SSL setting. On the other hand this information might be ambiguous, which could be resolved by using MIL. Thus, in the following, we combine both ideas and propose a new SSMIL approach, that integrates seamlessly information sources that are unreliable, ambiguous or both. Furthermore, in contrast to existing approaches, we can also integrate a multitude of different information sources.

We demonstrate our approach on a challenging dataset extracted from a TV series. However, the method is not limited to this scenario and can be easily adapted to other tasks. Succeeding we introduce the SSMIL algorithm. Further, we show in the experiments that a single weak information source suffices to obtain a reasonable performance gain over existing work.

2. SSMIL - Boosting

To make use of not fully reliable and ambiguous information cues, we address SSL and MIL in parallel. In particular, we realize this by formulating those concepts in a joint loss function, measuring the penalty for misclassifying training samples, and optimizing it using a Gradient Boosting framework. Thus, during learning we have to ensure that at least one sample of a positive bag is classified as positive whereas all instances of negative bags as negative. Additionally, the prior has to be approximated for the unlabeled bags.

2.1. Loss Function

Let $\mathcal{D}_l = \left\{(\mathcal{B}^l_1, y_1), \ldots, (\mathcal{B}^l_{N_l}, y_{N_l})\right\}$ and $\mathcal{D}_u = \left\{\mathcal{B}^u_1, \ldots, \mathcal{B}^u_{N_u}\right\}$ denote the set of labeled and unlabeled bags, 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 $y_i \in \mathcal{Y} = \{0, 1\}$ is the binary label for the respective bag. Further, we assume that a *prior* conditional probability $P_P(y|\mathcal{B})$ is given for the unlabeled bags. Then, the objective is to minimize the negative log-likelihood over both the labeled and unlabeled bags.

For the labeled bags, the loss can be written as

$$\mathcal{L}_l(\mathcal{D}_l) = -\sum_{i=1}^{N_l} \sum_{z \in \mathcal{Y}} [z = y_i] \log(P(y = z|\mathcal{B}^l_i)) \;, \quad (1)$$

where $[\cdot]$ is the Iverson bracket and $P(y|\mathcal{B}^l_i)$ is the bag posterior. Following the definition of MIL, the bag posterior is defined as

$$P(y = 1|\mathcal{B}^l_i) = \max_j P(y = 1|\mathbf{x}_{ij}) \;, \quad (2)$$

where $P(y = 1|\mathbf{x}_{ij})$ is the probability that an instance $\mathbf{x}_{ij}$ is positive.

For the unlabeled bags, following the approach of Saffari *et al*. [12], we define the loss over the unlabeled bags $\mathcal{L}_u$ as the deviation of the model from the prior. In detail, this is realized by measuring the cross entropy[1] between the prior and the model:

$$\mathcal{L}_u(\mathcal{D}_u) = -\sum_{i=1}^{N_u} \sum_{z \in \mathcal{Y}} P_P(y = z|\mathcal{B}^u_i) \log(P(y = z|\mathcal{B}^u_i)) \;. \quad (3)$$

Then, the overall loss function of the semi-supervised multiple instance problem can be written as the sum of both losses

$$\mathcal{L}(\mathcal{D}_l \cup \mathcal{D}_u) = \mathcal{L}_l(\mathcal{D}_l) + \lambda \mathcal{L}_u(\mathcal{D}_u) \;, \quad (4)$$

with $0 \leq \lambda \leq 1$ defining the influence of the unlabeled data.

2.2. Optimization – Boosting

To train a classifier, we need to optimize the loss function defined in Eq. (4). In general, any suitable optimization method could be applied. In particular, we propose to use Gradient Boosting [8].

[1] In the original formulation the Kullback-Leibler (KL) divergence between the priors and the bag posteriors is used. However, for the optimization problem the constant factors can be ignored simplifying the problem to the cross entropy.

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

$$F(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t f_t(\mathbf{x}) \ . \tag{5}$$

Thus, we can formulate our optimization problem as

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} \mathcal{L}(\mathcal{D}_l \cup \mathcal{D}_u) \ , \tag{6}$$

where the instance probability required for the loss function $\mathcal{L}$ is estimated by

$$P(y=1|\mathbf{x}) = \frac{e^{F(\mathbf{x})}}{e^{F(\mathbf{x})} + e^{-F(\mathbf{x})}} \ . \tag{7}$$

Update of Strong Classifiers Gradient Boosting then 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:

$$\frac{\partial \mathcal{L}(\mathcal{D}_l \cup \mathcal{D}_u)}{\partial F(\mathbf{x}_{ij})} \ . \tag{8}$$

Next, a new weak classifier $f_t(\mathbf{x})$ is constructed to produce outputs that approximate the inverse direction of this gradient (*i.e.*, reduce the residuals). The best weight α_t is then determined by a line search.

Update of Weak Classifiers In order to train weak learners in Gradient Boosting, we need to compute the partial derivatives of the loss function with respect to the response of the classifier to each instance. As can be seen in Eqs. (1) and (3), we therefore need the derivatives $a_{ij}(z)$ of the log-likelihood of the bags:

$$a_{ij}(z) = \frac{\partial \log P(y=z|\mathcal{B}_i)}{\partial F(\mathbf{x}_{ij})} \ . \tag{9}$$

From the definition of Multiple Instance Learning, the natural choice for the posterior probability of a bag being positive, would be Eq. (2). However, this measure is not differentiable, which is a prerequisite to use it within Gradient Boosting. To overcome this problem, the following approximations can be used:

Noisy OR Viola *et al*. [15]

$$P_{\text{NOR}}(y=1|\mathcal{B}_i) = 1 - \prod_{j=1}^{N_{B_i}} (1 - P(y=1|\mathbf{x}_{ij})) \tag{10}$$

Geometric Mean Lin *et al*. [9]

$$P_{\text{geo}}(y=1|\mathcal{B}_i) = 1 - \left[\prod_{j=1}^{N_{B_i}} (1 - P(y=1|\mathbf{x}_{ij})) \right]^{1/N_{B_i}} \tag{11}$$

Mean Pang *et al*. [10]

$$P_{\text{mean}}(y=1|\mathcal{B}_i) = \frac{1}{N_{B_i}} \sum_{j=1}^{N_{B_i}} P(y=1|\mathbf{x}_{ij}) \tag{12}$$

L_∞ Norm

$$P_{L_\infty}(y=1|\mathcal{B}_i) = \lim_{p\to\infty} \left(\sum_{j=1}^{N_{B_i}} P(y=1|\mathbf{x}_{ij})^p \right)^{1/p} \ . \tag{13}$$

In fact, the different posterior probabilities yield different estimates for $a_{ij}(z)$, however, the rest of the optimization problem is untouched.

Using $a_{ij}(z)$ we can derive the gradient of the loss function with respect to the output of the strong classifier $F(\mathbf{x}_{ij})$ as stated in Eq. (8). Thus, we can finally formulate the overall optimization process, yielding the t^{th} weak learner $f_t(\mathbf{x})$ as the dot product of the vector of all partial derivatives of the loss and the outputs of the new weak classifier:

$$\begin{aligned} f_t(\mathbf{x}) = \arg\max_{f(\mathbf{x})} & \sum_{i=1}^{N_l} \sum_{z\in\mathcal{Y}} [z=y_i] \sum_{j=1}^{N_{B_i^l}} a_{ij}(z) f(\mathbf{x}_{ij}) + \\ & + \lambda \sum_{i=1}^{N_u} \sum_{z\in\mathcal{Y}} P_P(y=z|B_i^u) \sum_{j=1}^{N_{B_i^u}} a_{ij}(z) f(\mathbf{x}_{ij}) \ . \end{aligned} \tag{14}$$

From this formulation we derive the following weight and label for each instance in order to train a weak classifier $f_t(\mathbf{x})$ optimizing Eq. (14)[2]. For instances in labeled bags, $\forall(\mathcal{B}_i^l, y_i) \in \mathcal{D}_l, \forall \mathbf{x}_{ij} \in \mathcal{B}_i^l$, we define the weights as

$$w_{ij} = |a_{ij}(y_i)| \tag{15}$$

and the labels are given by the bag label: $y_{ij} = y_i$. For instances in unlabeled bags, $\forall \mathcal{B}_i^u \in \mathcal{D}_u, \forall \mathbf{x}_{ij} \in \mathcal{B}_i^u$, the weights are defined by

$$w_{ij} = \lambda \left| \sum_{z\in\mathcal{Y}} P_P(y=z|\mathcal{B}_i^u) a_{ij}(z) \right| \ , \tag{16}$$

and *pseudo-labels* can be defined by looking at the direction of the gradient:

$$y_{ij} = \left[\sum_{z\in\mathcal{Y}} P_P(y=z|\mathcal{B}_i^u) a_{ij}(z) > 0 \right] \ . \tag{17}$$

[2]This way we could use any kind of weak learner. In our experiments we use decision stumps.

3. Face Recognition from Videos

In the following, we demonstrate our approach for learning face instance models from videos. In particular, we consider an episode from the TV series "Buffy the Vampire Slayer". The task is to assign a name to each face. The dataset of Everingham *et al*. [5] provides us with face tracks and appearance descriptors. In particular, face appearance is captured by a flexible part-based representation. Facial feature points are localized by a Pictorial Structures model [7]. The face descriptor is a concatenation of normalized pixel patches extracted at those locations. Further, face detections (in individual frames) are grouped into face tracks by motion information. Finally, a face track encodes the face appearance of a particular character within a shot.

3.1. Preparation of information cues

To augment the visual information, we exploit two main information sources closely associated to the video, namely transcript and subtitles; both containing the dialogs. Additionally, the transcript provides naming information and a textual description of what is happening; the subtitles set the dialogs into temporal context. To augment the transcript with the timing information it is aligned with the subtitles by dynamic time warping. From the transcript we infer the coarse scene structure, since it embraces scenes with the textual descriptions of what is happening. This is illustrated in Figure 2.

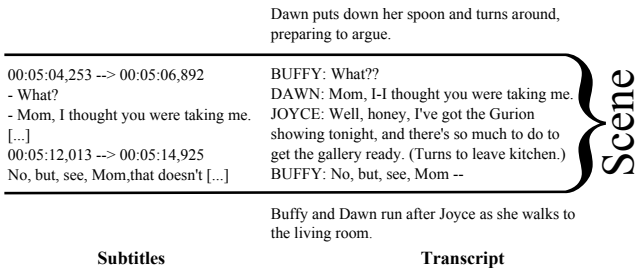


Figure 2: Coarse scene structure: The transcript[3] embraces scenes with textual descriptions of what is visually happening. With the augmented timing information of the subtitles these are put into temporal context.

In addition, the augmented transcript allows to infer the name of the speaker. Thus, we know who is speaking but neither if the speaker is visible nor to which face the text chunk belongs. We refine the candidate label by visual speaker detection. The decision if a face track is speaking or not is based on significant lip motion. For that purpose, we use the duality based TV-L1 method of [16]. Additionally, we estimate a global (head motion) motion compensation and finally just report the flow along the mouth normal (see Figure 3). Further cues like video editing rules [2], OCR [13], EPG or tags (dependent on the application scenario) could be used. Nevertheless, we show that in SSMIL a single weak additional information source is enough to achieve a significant performance gain over standard MIL. In the following we define the bags and introduce how we obtain the priors and labels.

[3] Obtained from the fan web-site http://www.buffyworld.com/. Subtitles are extracted of the DVD.

Figure 3: Visual speaker detection: Two succeeding face detections, (a) and (b) with localized facial features. The flow field in x direction (c) shows only minor, widespread motion. In contrast, the flow field in y (d) shows significant motion. In particular along the mouth normal.

3.2. Bag types

In order to encapsulate the available information cues we propose different bag types. A bag consists of one or more face tracks and an associated label or prior. The face tracks are represented by their individual face descriptions. Bags derived from the speaker detection are defined based on their creation rule. Intuitively, if a face track is detected as *speaking* we label it with the matching character name of the augmented transcript. Accordingly if a person is detected as *silent* we label it as negative for that cast name. If a face track is *coexistent* in time to a speaking one we label it as negative.

Furthermore, we define bags that contain all face tracks present in a scene, termed scene bags. The idea is to infer if a certain character is likely to appear in a particular scene or not. This is done dependent on the number of spoken text chunks. We empirically determine the probability that a character appears in a frontal pose in a temporal neighborhood around a subtitle. Then, the prior is approximated as binomial distribution, based on the number of subtitle appearances. One main benefit of the scene bags is that they capture some orthogonal information with respect to the visual speaker detection. For example, misses of the speaker detection and also *reaction shots* where in a dialog scene a character is only captured while not speaking. This is not possible in settings like [5, 6, 14].

3.3. Results

In the following, we evaluate our proposed method on the publicly available part of the *Buffy* dataset proposed by

Recall	50%	60%	70%	80%	90%	100%
P_{geo}	**89.2**%	82.7%	75.2%	69.6%	**66.1**%	**61.8**%
P_{L_∞}	88.5%	80.8%	75.8%	**70.4**%	65.2%	60.5%
P_{mean}	88.8%	**82.7**%	**75.8**%	69.9%	65.0%	61.4%
P_{NOR}	87.3%	80.8%	75.8%	70.1%	64.8%	60.9%

(a) MIL

Recall	50%	60%	70%	80%	90%	100%
P_{geo}	89.2%	85.3%	80.7%	76.1%	71.7%	66.9%
P_{L_∞}	**91.2**%	**86.9**%	80.2%	74.9%	70.6%	66.5%
P_{mean}	90.0%	86.5%	**82.9**%	**76.9**%	**72.3**%	**68.2**%
P_{NOR}	87.7%	85.9%	80.4%	76.1%	70.2%	64.9%

(b) SSMIL

Table 2: Precision values of the different models for the posterior probability of a bag. In the MIL case (a) the performance is quite similar. In contrast, for SSMIL (b) it is obvious that P_{mean} clearly outperforms the other bag posterior models.

Everingham *et al*. [5][4], which consists of 27504 individual frontal face detections. The task is to label each of the 516 face tracks by its cast name. The cast list of the ground truth annotation consists of 11 named entities, the class *other* and *false positive*.

For each cast member we train an one-vs.-all classifier. The training data contains the bags derived from the speaker detection and the scene bags. In total 259 tracks show persons when they are speaking. Using the combination of our visual speaker detection and the augmented transcript we label 173 tracks; 154 of those are correctly assigned. Detailed results of the labelling obtained by speaker detection for the individual cast members are reported in Table 1. 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. Precision means the ratio of correctly labeled samples, at the current threshold.

First we report the performances of the different models for the bag posterior probabilities on this task. The comparison is shown in Table 2. In the MIL case the performance of the different models is quite similar. In contrast, in the SSMIL case, especially for higher recall values, it is beneficial to use $P_{\text{mean}}(y|\mathcal{B}_i)$ as bag posterior model. Thus, in the succeeding experiment we use it as model for the bag posterior probability.

[4]The more recent "Buffy" dataset [14] is not publicly available.

In Figure 4 we benchmark our method with previous work [5, 6][5]. The baseline method classifies 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 follows:

$$d_f(F_i, F_j) = \min_{f_i \in F_i} \min_{f_j \in F_j} \|f_i - f_j\|, \qquad (18)$$

where $f_i \in F_i$ and $f_j \in F_j$ are face descriptions. According to [5] we also state the performance of labeling all face tracks with the cast name appearing most frequently in transcript (Prior on *Buffy*). Further, also the performance of using the aligned subtitles to propose a name is reported.

With the speaker 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 for the tracks detected as *speaking*. Therefore, the curve is constant for the first levels of recall. Due to the nearest neighbor classification the method has no real chance to recover from labeling errors. Label noise propagates directly into the classification. If the method labels all face tracks a precision of 60.1% is reached. Already MIL outperforms the baseline over most levels of recall – at 100% recall the precision is 61.4%. SSMIL, however, yields a clear additional improvement. At 100% recall we obtain a precision of 68.2%, an improvement of 8.1% over the baseline. Indeed, the method even delivers a higher precision as the speaker detection up to a recall level of 54%. It labels 20% more tracks with the same accuracy of 89%. This shows the ability of SSMIL to recover from labeling errors.

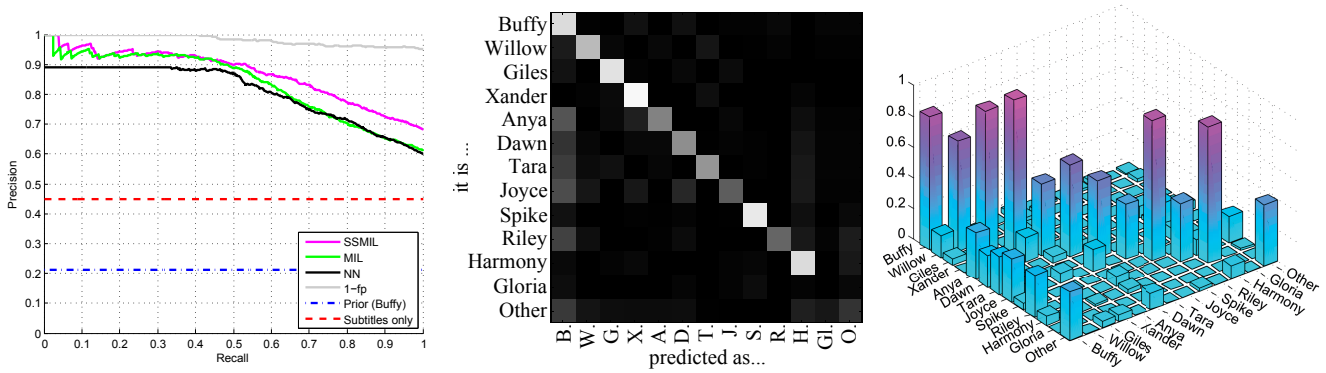
4. Conclusion

In this work we presented the task of face recognition in weakly labeled videos as Semi-Supervised Multiple Instance Learning (SSMIL) problem. Multiple Instance Learning enabled us to incorporate ambiguous information that relates to a bag of instances. Semi-Supervised Learning allowed us to make use of not fully reliable information. By formulating those concepts in a joint loss function, that measures the penalty for misclassifying training samples, we are able to optimize it in a Gradient Boosting framework. Gradient Boosting allows to use any suitable loss function as long as it is differentiable. To demonstrate the strength of our method, we evaluated it on the publicly available part of the *Buffy* dataset, comparing to the baseline method proposed by [5]. Already MIL outperformed the baseline over most levels of recall. Moreover, SSMIL revealed a further clear improvement. In particular, we showed that for SSMIL only one additional weak information cue suffices to improve the performance over the

[5]Unfortunately it is not possible to directly compare to the numbers of the original publication [5, 6] as some important data (speaker detection, clothing descriptors) is not provided.

Name	*Buffy*	*Willow*	*Giles*	*Xander*	*Anya*	*Dawn*	*Tara*	*Joyce*	*Spike*	*Riley*	*Harm.*	*Other*
#tracks	110	42	24	30	13	72	16	14	22	29	71	48
TP	39	14	7	16	4	8	5	3	10	5	37	6
FP	4	2	2	2	1	0	0	1	0	2	1	4
Precision	0.91	0.88	0.78	0.89	0.80	1.00	1.00	0.75	1.00	0.71	0.97	0.60
Recall	0.64	0.64	0.50	0.80	0.57	0.44	0.50	0.30	0.59	0.45	0.67	0.43
Coverage	0.35	0.33	0.29	0.53	0.31	0.11	0.31	0.21	0.45	0.17	0.52	0.13

Table 1: Performance analysis of the labeling obtained by the visual speaker detection: *TP* and *FP* denote the number of true positives and false positives respectively. While *Recall* refers to the set of all *speaking* tracks, *Coverage* means the percentage of correctly labeled tracks in relation to the total number cast appearances (including *non-speaking* tracks).



(a) Precision/Recall

(b) Confusion Matrix

Figure 4: Buffy dataset. (a) SSMIL clearly outperforms the baseline (NN) over all levels of recall. (b) The associated confusion matrix.

baseline. As we are not limited to specific cues or features the method is easily extendable. For instance it is possible to incorporate other appearance descriptors or bag types.

References

[1] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *Proc. CVPR*, 2009.

[2] T. Cour, B. Sapp, A. Nagle, and B. Taskar. Talking pictures: Temporal grouping and dialog-supervised person recognition. In *Proc. CVPR*, 2010.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[4] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997.

[5] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *Proc. BMVC*, 2006.

[6] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5), 2009.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Intern. Journal of Computer Vision*, 61:55–79, 2005.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–374, 2000.

[9] Z. Lin, G. Hua, and L. Davis. Multiple instance feature for robust part-based object detection. In *Proc. CVPR*, 2009.

[10] J. Pang, Q. Huang, and S. Jiang. Multiple instance boost using graph embedding based decision stump for pedestrian detection. In *Proc. ECCV*. 2008.

[11] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *Proc. ICCV*, 2007.

[12] A. Saffari, H. Grabner, and H. Bischof. SERBoost: Semi-supervised boosting with expectation regularization. In *Proc. ECCV*. 2008.

[13] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh. Video OCR: indexing digital news libraries by recognition of superimposed captions. *Multimedia Systems*, 7:385–395, 1999.

[14] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" Learning person specific classifiers from video. In *Proc. CVPR*, 2009.

[15] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *Advances in NIPS*. 2006.

[16] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *Proc. BMVC*, 2009.