

# Time-aware Co-Training for Indoors Localization in Visual Lifelogs \*

Vladislavs Dovgalecs

Rémi Mégret

Yannick Berthoumieu

University of Bordeaux  
IMS, UMR 5218 CNRS  
F33400 Talence, France

{vladislavs.dovgalecs, remi.megret, yannick.berthoumieu}@ims-bordeaux.fr

## ABSTRACT

In this paper we address the problem of location recognition from visual lifelogs by leveraging visual features and temporal information in a unified framework. The proposed method features a co-training approach that takes advantage of both labeled and unlabeled data using a confidence measure we propose for this task. It exploits jointly two SVM classifiers on two types of visual features as well as the temporal continuity of the video through temporal accumulation scheme. We demonstrate experimentally on the publicly available IDOL2 dataset that the algorithm yields performance improvement due to its ability to exploit jointly multiple cues, time and unlabeled data.

## Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – Indexing methods; I.5.2 [Pattern Recognition]: Design Methodology – Classifier design and evaluation, Feature evaluation and selection.

## General Terms

Algorithms, Design, Experimentation.

## Keywords

Lifelog indexing, semi-supervised learning, co-training, temporal information.

## 1. INTRODUCTION

Lifelogging and ego-centric video monitoring is now a practical way for activity and behavior monitoring [7] as well as a mean for memory aid. Although there exist technical

\*Area Chair: Kiyoharu Aizawa

Acknowledgment: This work is partially supported by a grant from Agence Nationale de la Recherche with reference ANR-09-BLAN-0165-02, within the IMMED project <http://immed.labri.fr/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...\$10.00.

solutions for activity log acquisition, the need for efficient content structuring is still of great importance and remains an open question. To search multiple hour recordings efficiently, an indexing method is required. In our study we focus on the estimation of the location of the monitored person using visual information captured by a wearable wide angle camera attached to the person's shoulder.

Indoors localization using image-based approaches is a widely researched topic in computer vision and robotics. Discrete features [12] can be used for indoors and outdoors localization but require explicit matching of local features which can be costly. Signature based approaches such as Bag of Features (BOF) [8] or CRFH [5], which are based on the distribution of quantized local features, produce global image description that is more easily amenable to efficient use for video analysis.

An important issue concerns the discrimination power. A single type of features may not be efficient in all conditions. Fusion of multiple cues is a way to improve the discriminability by exploiting several feature types that exhibit complementary properties with respect to selectivity and invariance. A multiple cue integration method exploiting confidence measure for place classification was proposed in [9], which uses the late fusion of SVM scores obtained on diverse visual cues. They also propose temporal accumulation of the scores over time, which allows to smooth out isolated misclassifications.

Another specificity of lifelog indexing is the low amount of training data. Semi-supervised learning is a way to profit from abundant unlabeled data together with small labeled data set. It was used for instance in [3] to improve BOF and feature matching recognition in the case of sparse labeling.

In this paper we investigate how to combine both semi-supervised, temporal and multiple visual cues for location recognition. We show that appearance-based model can be refined iteratively in a semi-supervised manner using predictions and a confidence measure. Secondly, we show how two cues can be fused to train a classifier in a co-training setup for image-based localization while taking into account both unlabeled data and temporal information.

The paper is organized as follows. In section 2, we introduce a supervised approach featuring multiple cue fusion and temporal accumulation that is representative of the state of the art and will serve as a baseline. In section 3, we propose semi-supervised time-aware co-training method. Results are presented in section 4.

## 2. SUPERVISED FUSION OF MULTIPLE CUES AND TIME INFORMATION

In this section we review a multiple cue fusion method proposed in [9]. We complement it with a non-linear dimensionality reduction step, which allows the rest of the framework to work on low dimensional features with linear SVM kernel, and temporal accumulation. This method is used as a final stage module in the proposed framework.

### 2.1 Feature extraction and preparation

For each image of the video we extract two visual features: BOF [8] (hierarchical vocabulary with 1111 words) and very high dimensional but sparse CRFH [5] gist-like features. The intersection kernel [1] is used to evaluate affinities between such distribution based features. Other image kernels such as local feature matching [12] are not considered here because of their higher computational cost, but could also be used.

The extracted visual features are of high dimensionality which poses a problem of over-fitting when the number of training samples is low. We reduce the dimensionality using KPCA [11] to 100 dimensions which decreases the risk of over-fitting. The selected intersection kernel also takes into account the non-linearity of the original feature spaces, yielding a linear embedding space that approximates the original affinities.

Additionally, we noticed that affinity matrix pruning help improving the results. A simple k-Nearest Neighbor pruning decreased harmful influence of distant and non-similar images. Computation of linear kernel from KPCA embeddings ensures positive definite kernel for linear kernel classifier.

In the rest of this paper, classification will be done using linear SVMs applied on the KPCA embeddings.

### 2.2 SVM-based Discriminative Accumulation

Let  $\mathbf{x}$  be a pattern (an embedding vector obtained from the preprocessing step) and  $y \in \{-1, 1\}$  the corresponding label in a two-class context. Given a test pattern  $\mathbf{x}$ , the SVM learns a decision function

$$f(\mathbf{x}) = \sum_{i=1}^n y_i \alpha_i k(\mathbf{x}_n, \mathbf{x}) + b$$

where we selected a linear kernel  $k(\mathbf{x}_n, \mathbf{x}) = \mathbf{x}_n^T \mathbf{x}$ . The patterns  $\{\mathbf{x}\}_{i=1}^n$  and associated  $\{\alpha\}_{i=1}^n$  are the training patterns and learned coefficients.

For  $C > 2$  classes, we train  $C$  functions  $\{f_j\}_{j=1}^C$  in a one-versus-all setup. For a test pattern  $\mathbf{x}_i$  we therefore obtain  $C$  decision values or scores  $s_j^i = f_j(\mathbf{x}_i)$ ,  $j = 1, \dots, C$ . The class estimation is obtained as

$$\hat{y}_i = \arg \max_{j=1..C} s_j^i$$

The Discriminative Accumulation Scheme (DAS) proposed in [9] improves the performance by fusing the information from two different classifiers in a late fusion approach: the new scores are built by combining linearly the two independent SVM classifier outputs

$$s_{j,DAS}^i = \beta s_{j,BOF}^i + (1 - \beta) s_{j,CRFH}^i$$

As pointed out by the authors and reference in [9], the accumulation scheme makes the decision more robust compared to the majority vote approach.

### 2.3 Temporal Accumulation Scheme

Additionally, images from the video that are close in time are likely to belong to the same class. This can be taken into account by using Temporal Accumulation (TA)

$$s_{j,TA}^i = \sum_{\Delta i = -\tau}^{\tau} w(\Delta i) s_j^{i+\Delta i}$$

where  $w(\Delta i)$  represents the weight of a temporal window of size  $2\tau + 1$ . Effectively, large score variation for neighbor images is lessened by this operation of smoothing.

## 3. SEMI-SUPERVISED TIME-AWARE CO-TRAINING APPROACH

We now extend the previously introduced supervised framework to a semi-supervised approach using the co-training paradigm.

### 3.1 Co-training Architecture

The idea of self-training [10] consists in supplying the most confident estimations to the labeled set where the procedure can be repeated on the remaining unlabeled patterns. This type of learning can be extended to two visual cues, giving rise to co-training [2] where each single classifier trains another one with the most confident estimates. For a single view, [4] proposed a method which trains two learners by using two different learning algorithms. Tri-training [14] teaches a third classifier the consensus obtained of two other classifiers. The co-training approach is interesting for our application as it brings together the advantages of multi-cue fusion with semi-supervised learning [13].

We now review the co-training algorithm which iteratively learns two SVM classifiers. The method is part of architecture shown in Fig. 1.

At round  $t = 0$  two SVM classifiers are learned independently on the training set in a one-against-all setup. Initial estimations are produced on the unlabeled data set. Applying the DAS method on these two outputs represent a baseline supervised method.

At round  $t = n$  the  $p$  most confident estimates from each view are selected and appended to the opposite training set. The concerned patterns are then removed from both testing sets and a new pair of classifiers is learned after this feedback loop. The procedure may be repeated until exhaustion of unlabeled patterns. Applying the DAS method on the SVM scores produced by these new classifiers corresponds to what we call the co-training-DAS approach (CO-DAS).

In the experimental section we will consider only one iteration of CO-DAS in order to investigate the gain of the proposed feedback loop compared to other approaches.

The performance of the approach is highly dependant on the selection of the most confident estimates to be used in the feedback loop. Ideally, only correct estimates should be used. Finally, complementary visual cues will allow to learn an enriched model in order to avoid over-fitting. For this task, in the following discussion, we contribute a confidence measure, and introduce the framework incorporating time information which provides more diverse training patterns into the learning loop.

### 3.2 Selection of confident predictions

The proposed selection scheme exploits the SVM scores to construct a scalar confidence measure  $z_i = z(\mathbf{x}_i)$ . It is

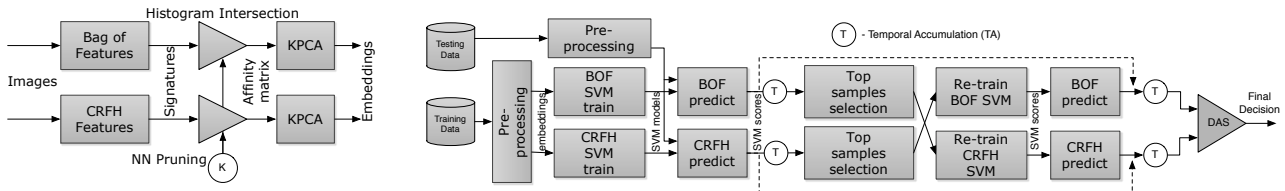


Figure 1: (Left) Pre-processing for embedding computation. (Right) Time-aware co-training workflow

designed to be positive and larger for confident and close to zero for less confident estimates. Our confidence measure  $z_i$  is computed in two steps. First the contrast between the best score and the rest is computed

$$z_i^0 = s_{j^*}^i - \sum_{j \neq j^*} s_j^i$$

It is then penalized in case of class overlap

$$z_i = z_i^0 \max\left(0, 1 - \frac{p_i - 1}{C}\right)$$

where  $p_i = \text{Card}(\{j = 1..C | s_j^i > 0\})$  represents the number of classes for which  $\mathbf{x}_i$  has positive scores. In the degenerate case of all positive or negative scores values, we set zero confidence ( $z_i \leftarrow 0$ ).

### 3.3 Time-aware feedback

Time information can be injected using the Temporal Accumulation Scheme (TA) proposed in section 2.3 prior the final decision. The main contribution of this paper shows that temporal information should be used in the co-training loop. It is done by applying TA on the scores in the inner feedback loop, thus defining a new time-aware co-training approach. One of the main advantages is that TA is not a mere post-processing, but can help bringing new patterns that are visually different from the training samples, provided they are temporally close to confident samples. More diversity is introduced, thus yielding models that generalize better. The temporal window should not be selected too large, in order to avoid accumulating over the temporal boundaries between classes. The experiments presented in the next section confirm the interest of this approach.

## 4. EXPERIMENTAL PERFORMANCES

In this section we report the performances obtained on public real-world data to evaluate the gain brought by the proposed approach.

### 4.1 Considered approaches

All approaches compared in this section can be defined from a set of building blocks that have been defined previously.

We define three main categories. In each of them the TA variant appends Temporal Accumulation as a post-processing before taking final decision on the class. For the sake of readability, only the DAS decision approach is considered here, as it provides consistently better results than classifying the BOF or CRFH features separately.

- DAS, TA-DAS: supervised approaches;
- CO-DAS, CO-TA-DAS: semi-supervised approaches; the confidence values used in co-training are based on the raw scores from the individual classifiers;

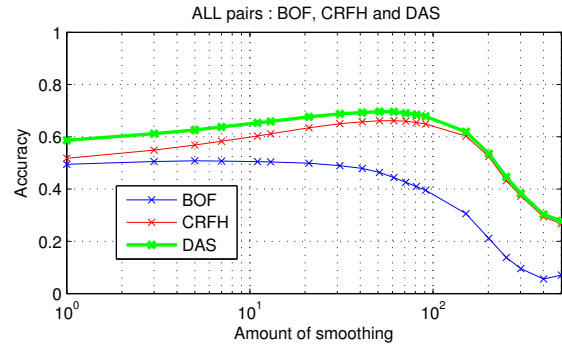


Figure 2: Effect of both multi-cue fusion and temporal accumulation on supervised classifier

- TA-CO-DAS, TA-CO-TA-DAS: semi-supervised time-aware approaches; the confidence values used in co-training are based on temporally accumulated scores.

### 4.2 Data corpus

In our experiments we used the IDOL2 video database [6] which is representative of the appearance of a camera moved across an indoor environment. It includes visual variabilities due to different lighting conditions (sunny, cloudy, night) and natural change of scene due to human activities over time. We used the part called “minnie”, which contains 12 individual sequences captured at 5 fps, for a total duration of 38 minutes, representing repeated visits through 5 different rooms at various times and light conditions. More details on this corpus can be found in [6].

All possible pairs of sequences have been tested, using one sequence for training and one for testing. 12 pairs represent same light conditions with close in time recording, 48 pairs represent different light conditions with several weeks between the recording. Average performances are considered.

### 4.3 Effect of temporal smoothing

In Fig. 2 the effect of both multi-cue fusion using DAS and Temporal Accumulation is shown.

The results show that the TA scheme yields an improvement over the DAS baseline by around 10% where no smoothing corresponds to the leftmost value (temporal window of 1 sample). This is true in cases of different levels of annotations, for different light conditions and sequence capture times. The optimal values is around 50 frames which is not surprising if we take into account the fact that the robot was constantly moving and did not spend more than 10 seconds in the same room. Larger temporal window may result in severe misclassifications around class boundaries. For subsequent tests we selected a temporal window of 50 frames.

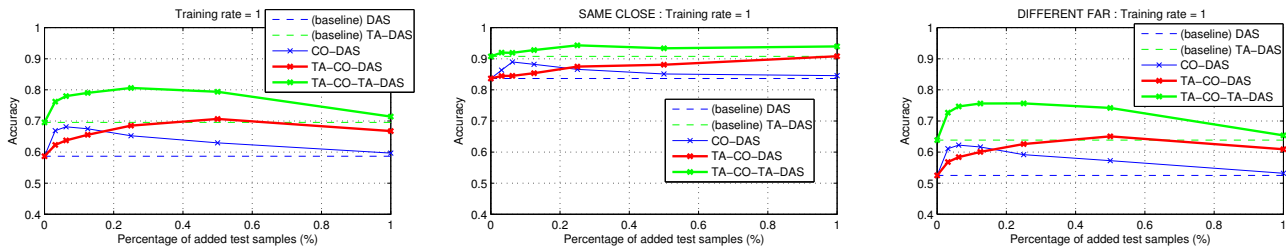


Figure 3: Performance of co-training and its interaction with temporal accumulation schemes.

#### 4.4 Effect of co-training

In Fig. 3 the performance of the proposed approach is shown, as well as a detailed comparison to other approaches including temporal information. The figure shows a summary of average performance for all possible sequence pairs (a), as well as the average performance for the most extreme cases: same light, close in time sequence pairs (b) and different light, far in time sequence pairs (c). In each summary the baseline DAS scheme is compared to the TA-DAS baseline and three extended methods involving co-training, temporal accumulation and DAS scheme in different setups. We plot the average performance for each method as a function of the amount of training samples used for the co-training feedback loop.

First, we can note strikingly different performances due to different light conditions and complexity of the scenes caused by a difference in time for certain sequence pairs. Nevertheless, the relative ranking of the studied approaches remains the same in both extreme cases, which is why we will discuss only the global average.

Indeed, the TA-CO-DAS and TA-CO-TA-DAS methods show a consistent improvement compared to their respective baselines (DAS and TA-DAS) not using co-training.

It is of interest to note that co-training without any temporal information (CO-DAS) improves on its baseline, reaching almost the level of TA-DAS. Nevertheless, temporal information is best used both in the co-training loop and at post-processing step yielding TA-CO-TA-DAS method. This yields the best performances, showing the ability of the proposed method to include all considered sources of information.

## 5. CONCLUSION

In this paper, we have presented a semi-supervised approach within the co-training framework for video lifelog indexing which profits both from discriminative kernel learning methods and from the complementarity of two visual features integrated in a time-aware semi-supervised framework.

The proposed approach, by selecting the most reliable samples for the iterative co-training steps, allows us to adapt the model initially trained to the characteristics of the unlabelled data, thus improving the performances. Based on our experiments on the IDOL2 database, the improvement in performance is larger when time information is taken into account, both during co-training feedback and as a post-processing. This confirms that the semi-supervised closed-loop brings relevant information, that could not be extracted in the supervised approach, and that the use of time information is a central part of this improvement.

## 6. REFERENCES

- [1] A. Barla, F. Odone, and A. Verri. Histogram Intersection Kernel for Image Classification. *Int. Conf. on Image Processing (ICIP)*, 2003.
- [2] A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. *Int. Conf. on Learning Theory (COLT)*, 1998.
- [3] V. Dovgalecs, R. Megret, H. Wannous, and Y. Berthoumiu. Semi-Supervised Learning for Location Recognition from Wearable Video. *Content-Based Multimedia Indexing (CBMI)*, 2010.
- [4] S. Goldman and Y. Zhou. Enhancing Supervised Learning with Unlabeled Data. *Int. Conf. on Machine Learning (ICML)*, 2000.
- [5] O. Linde and T. Lindeberg. Object Recognition using Composed Receptive Field Histograms of Higher Dimensionality. *Int. Conf. on Pattern Recognition (ICPR)*, 2004.
- [6] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt. The KTH-IDOL2 Database. Technical report, Kungliga Tekniska Högskolan, CVAP/CAS, 2006.
- [7] R. Megret, V. Dovgalecs, H. Wannous, S. Karaman, J. Benois-Pineau, E. El Khory, J. Pinquier, P. Joly, R. Andre-Obrecht, Y. Gaestel, and J.-F. Dartigues. The IMMED Project: Wearable Video Monitoring of People with Age Dementia. *ACM Multimedia*, 2010.
- [8] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.
- [9] A. Pronobis and B. Caputo. Confidence-Based Cue Integration for Visual Place Recognition. *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2007.
- [10] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-Supervised Self-Training of Object Detection Models. *7th IEEE Workshop on Applications of Computer Vision (WACV)*, 2005.
- [11] A. J. Smola, B. Scholkopf, and K.-R. Muller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural computation*, 10(6):1299–1319, 1998.
- [12] C. Valgren and A. J. Lilienthal. SIFT, SURF & Seasons: Appearance-Based Long-Term Localization in Outdoor Environments. *Robotics and Autonomous Systems*, 58(2):149–156, 2010.
- [13] Z. H. Zhou. When Semi-Supervised Learning meets Ensemble Learning. *Workshop on Multiple Classifier Systems (MCS)*, 2009.
- [14] Z. H. Zhou and M. Li. Tri-training : Exploiting Unlabelled Data using Three Classifiers. *IEEE Trans. on KDE*, 17(11):1529–1541, 2005.