A DICTIONARY LEARNING APPROACH TO TRACKING

Mark Barnard, Wenwu Wang, Josef Kittler

The Centre for Vision
Speech and Signal Processing
University of Surrey
Guildford GU2 7XH, UK.
{mark.barnard, w.wang, j.kittler}@surrey.ac.uk

Syed Mohsen Naqvi, Jonathon A. Chambers

Advanced Signal Processing Group
School of Electronic and Elec. and Sys. Eng.
Loughborough University
Loughborough, Leicester, UK.
{s.m.r.naqvi, j.a.chambers}@lboro.ac.uk

ABSTRACT

The problem of tracking people using multiple cameras is of much current interest as a means of providing cues for audio-visual blind source separation in dynamic environments. Here we investigate the use of one of the current state-of-the-art techniques in object recognition combined with one of the most popular methods of modelling object motion, particle filters, for tracking people. The dictionary learning or Bag-of-Words approach to object recognition has proved to be very effective in recent years, as shown in a number of large comparisons such as the PASCAL Visual Object recognition Challenge (VOC). In this paper we use this proven object recognition method within the framework of a particle filter. This provides a more accurate and robust tracking of people in a multiple camera environment. We also demonstrate that the dictionary learning approach can provide a principled method for the fusion of multiple features.

Index Terms— Tracking, Dictionary Learning

1. INTRODUCTION

The tracking of people has attracted significant interest from researchers in the area of blind source separation as an additional cue for audio-visual source localisation [1]. The problem of visual tracking can be decomposed into three sub-problems [2]: image representation, modelling the appearance of the object and modelling the motion of the object. In terms of tracking appearance modelling presents the greatest challenge as the appearance of the object can change greatly over time. These problems generally require efficient processing of large amounts of data in real time or near real time. Due to this constraint many of the current state-of-the-art approaches to object recognition have not been considered when it comes to tracking. The recent increase in the available processing power and storage space is however now making it possible to use these approaches not previously considered for tracking.

If we consider appearance modelling as essentially an object recognition task then there are two options for modelling the variability in object appearance, either pre-training the models using large amounts of training data or updating the models online as the object's appearance changes. The disadvantage of the first approach is that the training set may not be comprehensive enough to include all the variation in the object's appearance. The disadvantage of the second approach is that any error in the tracking will cause an error in the model as it is updated. This leads to the tracker drifting from the desired object and eventually failing.

The objective of this paper is to present one of the current state-of-the-art approaches in the area of visual object recognition, namely dictionary learning or the so called bag-of-words/features, within the framework of the well established tracking algorithm, particle filters [3]. This combines the discriminative power of dictionary learning with the localisation and dynamic modelling abilities of particle filters. We demonstrate the effectiveness of this method in tracking people in a meeting room environment using multiple cameras.

It has been shown in previous work [4, 5] that the use of multiple modalities can improve robustness in tracking applications. In our case we use colour histogram features to model the colours of the face and Scale Invariant Feature Transform (SIFT) descriptors [6] to capture the shape and texture of the face. We incorporate these two features in a principled way by creating a single dictionary of the combined features. The use of a discriminative classifier, a Support Vector Machine (SVM), enables the modelling of not just the object we wish to track but also the background.

Object detection and appearance modelling can often be considered as parts of the same problem, that of recognising an object under many different appearance conditions. There are broadly two approaches to this problem based on either static appearance models or adaptive appearance models. Many early approaches to tracking used an initial appearance to either manually define or train a model of the object [7, 8].

This research was supported by the Engineering and Physical Sciences Research Council of the UK (grant no. EP/H050000/1).

These static appearance models have difficulty in coping with changes in object appearance as tracking continues.

The problem of modelling object appearance with static models can be overcome by having enough training data to capture the variability in the object's appearance. However collecting large enough amounts of training data can be prohibitively expensive. Özuysal et al. [9] present a system that uses a sequence of a slowly moving object as the training set for a set of randomised trees. Features are selected based on the number of times they are successfully detected to prevent the set of features from becoming prohibitively large. An initial small set of fully labelled data, combined with larger sets of weakly labelled and unlabelled data, is used in [10] with semi-supervised learning to generate an appearance model.

An alternative approach to modelling object appearance over time is to adapt object models online. Ross et al. [11] incrementally learn a low-dimensional subspace representation of the object being tracked. Multiple Instance Learning is used to learn positive and negative examples of the object being tracked by Babenko et al. [2] in order to adapt the appearance model for face tracking. A particle filter is used by Wang et al. [12] with Haar wavelet features to select object and background regions and use these to update the appearance models online.

In the following section we present a system that combines a powerful method of visual object recognition and appearance modelling, based on dictionary learning, with the current state-of-the-art method for motion modelling, using particle filters. In Section III we describe the experiments conducted and the data used together with the results obtained. Conclusions are given in Section IV.

2. OUR APPROACH

In this section we present a system for tracking using a pre-trained dictionary and SVM classifier within the framework of a particle filter to provide robust and accurate 3D tracking using multiple cameras. We take the approach of pre-training a model of the object to be tracked, in this case a person's face and head. We believe that given a large enough and rich enough set of training data we can model a wide range of object appearance. This will overcome the problem of errors compounding that can occur when models are updated online. Using the method described in the following section we can quickly and efficiently generate large amounts of varied training data. We apply the method to generate approximately 28,000 training examples, for both object and background.

2.1. Dictionary building

We build a dictionary based on the features obtained from image patches extracted from the training data by the method discussed in the previous section. We extract two types of features to characterise each patch: SIFT descriptors, 128 dimensions and Hue histograms, 100 dimensions. We create individual dictionaries for each type of descriptor and also a dictionary using a concatenation of the two features.

We define a dictionary as a set $V = \{X_0, X_1 \dots X_{N-1}\}$ where $X_i, i = 0, \dots, N$ is a vector of features. All descriptors in the training set are clustered using the K-means algorithm into N clusters with the centre of each cluster being an atom in the dictionary. As there is no previous work using dictionary learning for this application we decided to test a number of different values for N. We need to balance the number of atoms in the dictionary between being able to discriminate the object we wish to track and over-fitting on the training set.

We used a hierarchical k-means process. First the data are clustered into N_h high level clusters and then N_l lower level clusters where $N = N_h \times N_l$. This allows us to efficiently cluster large amounts of training examples. A coefficient vector is then produced for each patch in the training set based on the atoms in the dictionary. The coefficient vector is an N bin histogram which is populated using a soft voting technique. Soft voting uses the codeword uncertainty method presented in [13] where the entry in the coefficient vector C for each atom w is given by

$$C(w) = \frac{1}{t} \sum_{i=1}^{t} \frac{K_\sigma(D(w, r_i))}{\sum_{j=1}^{N} K_\sigma(D(w_j, r_i))}, \quad (1)$$

where t is the number of descriptors in the image, $D(w, r_i)$ is the Euclidean distance between dictionary atom w and the descriptor r_i, K is a Gaussian kernel with smoothing factor σ and V is the dictionary containing the atom w. The variance of the Gaussian kernel K is given by σ^2 and so the value of σ can be seen as a control on the sparsity of the coefficient vector C.

In the initial presentation of this method the authors estimated the value of the smoothing factor σ experimentally using a training and validation set. In our case we estimated σ directly from the data by taking the standard deviation of the distribution of distances from descriptors to their cluster centres. This method proved to be much faster while still producing a reasonable estimate of σ. The *Codeword Uncertainty* method of histogram generation, shown in Equation 1, has been shown to perform well in the PASCAL VOC challenge [14].

2.2. Tracking system

Using the dictionary obtained above we create coefficient vectors for all the object and background training patches. Now an SVM is trained using these labelled coefficient vectors as shown in Figure 1. The trained SVM is then used within the framework of a particle filter in order to track people's heads. To model the object dynamics we used a first order particle filter with the state space composed of the objects position,

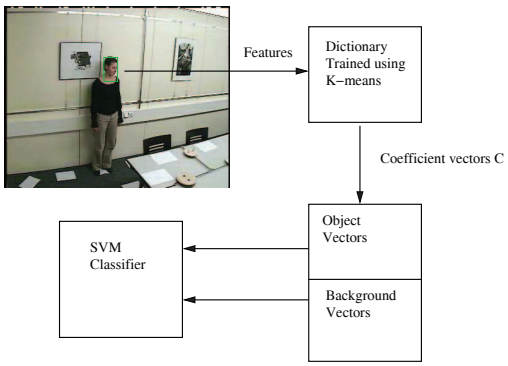

Fig. 1. Training an SVM classifier using labelled coefficient vectors generated from a dictionary.

scale and velocity. The number of particles was set to 50 for all experiments. The weight of each particle is given by the likelihood of the object versus background produced by the pre-trained SVM. An outline of the measurement step of the particle filter used in a tracking mode is shown in Figure 2. For each particle SIFT and Hue features are extracted and the dictionary is used to create a coefficient vector which is then classified by the SVM. The likelihood from the SVM is used to determine the weight of the particle.

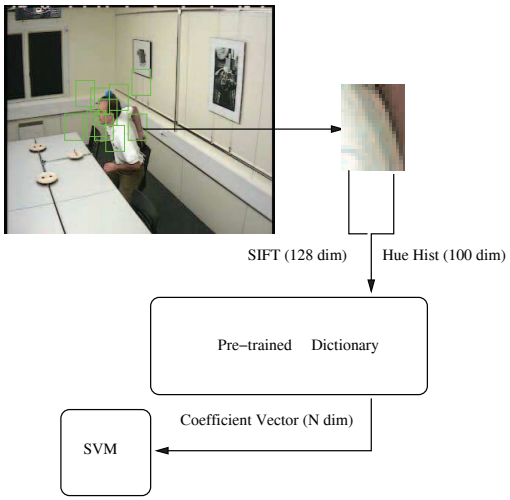

Fig. 2. The measurment step of the proposed tracking system.

3. EXPERIMENTS

We created a dictionary and trained an SVM as described in Section 2 using training data collected from sequences of the AV16.3 database [15]. We then tested this model using three more difficult sequences from the same database. Each sequence is between 1000 and 1500 frames long with a frame rate of 25 frames per second and each video frame is a colour image of 288x360 pixels.

Initial tests were conducted with a dictionary size of 64 atoms and separate dictionaries for Hue and SIFT. After these initial experiments a number of different dictionary sizes containing respectively 32, 64, 128, 256, 512, and 1024 atoms were tested. Although in visual object recognition tasks larger dictionary sizes are commonly adopted we decided for practical purposes to limit the maximum size to 1024 atoms.

3.1. Training data collection

The data used in our experiments consists of 6 annotated and 12 unannotated sequences, recorded in a meeting room environment using 3 calibrated cameras. These feature a single subject moving within the field of view. The sequences vary in difficulty from the subject simply moving around a set of positions in the room with relatively constant direction and velocity, to the subject moving freely around the room and making abrupt changes in direction. From the annotated sequences we selected 3 for training and 3 for testing. The variability of appearance in the training data was maximised by combining data from all three cameras to train a single model.

In order to collect training data for both the object to be tracked and the background we used a semi-supervised version of a baseline tracker. This is a simple particle filter using Hue histogram distance as the measurement step. An initial exemplar patch of the face is taken for each camera and the Battacharya distance is then taken for each particle to determine it's weight. This method is effective for tracking simple sequences and can be re-initialised by hand when it does fail. We found this to be an effective method of collecting large amounts of object training data. The background data was collected using frames with no subject in them and frames from each sequence were used to take into account small changes in the room layout and lighting conditions.

3.2. Results

We measured the performance by measuring the Mean Squared Error between the 3D position given by each method and a manually annotated 3D position. The results for single features Hue histogram and SIFT are shown in Table 1. Here we compare the approach based on histogram distance, as described in the previous section, with our proposed dictionary learning approach. It can be seen that for all sequences dictionary learning outperforms the baseline approach. In sequence 3 the performance of the Hue mode is very poor, this can be explained by Figure 3 where the Hue tracker has failed completely as the subject turns his head. The Hue histogram and Hue dictionary methods are both recognising the skin colour on the arm instead of the head. In order to improve robustness it was decided to fuse the Hue histogram and the SIFT features. The results of this fusion with various dictionary sizes can be seen in Figure 4. In all cases the fused

feature dictionary outperforms the single feature dictionary. Interestingly the best performance comes from the smallest dictionary size, this could be due to larger dictionary sizes over-fitting.

Sequence	Hue hist	SIFT	Hue dict	SIFT dict
Sequence 1	10.90	12.64	9.16	9.75
Sequence 2	13.03	15.16	9.59	10.00
Sequence 3	22.36	13.42	15.16	10.48

Table 1. The results for tracking shows the Root Mean Squared Error in centimetres between each test sequence using SIFT and Hue histogram

Fig. 3. The left image shows the results using Hue and the right image shows the combined Hue and SIFT dictionary approach. The red circle shows the tracking result.

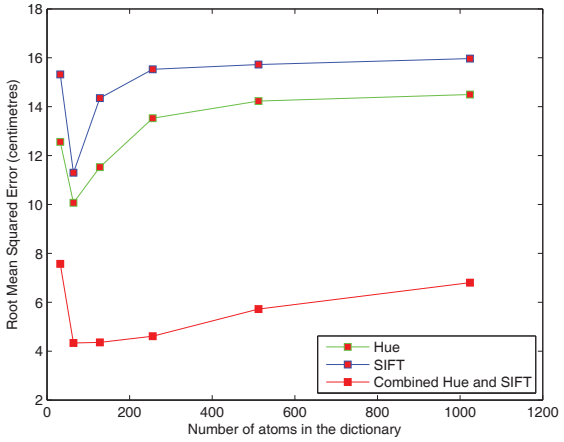

Fig. 4. Plot of Root Mean Squared Error in centimetres for Hue, SIFT and combined Hue and SIFT for varies dictionary sizes.

4. CONCLUSIONS

We have proposed a tracking system combining dictionary learning for object recognition and appearance modelling with a particle filter for motion modelling. This proposed method was shown to be more accurate than baseline methods on a challenging person tracking database. We also demonstrated that the combination of Hue and SIFT descriptors with a dictionary learning framework provides a more robust tracker. Whilst it can be argued that this performance improvement comes at the expense of model complexity, we believe that improvements in computing power will make these approaches more practical in the future.

5. REFERENCES

[1] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, 2010.

[2] B. Babenko, M.-H. Yang, and S. Belongie, "Visual Tracking with Online Multiple Instance Learning," in *CVPR*, 2009.

[3] M. Isard and A. Blake, "Condensation conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[4] T. Chang, S. Gong, and E. Ong, "Tracking multiple people under occlusion using multiple cameras," in *Proceedings of British Machine Vision Conference*, 2000.

[5] F. Talantzis, A. Pnevmatikakis, and L. C. Polymenakos, "Real time audio-visual person tracking," in *Proceedings of IEEE 8th Workshop on Multimedia Signal Processing*, 2006.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[7] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[8] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[9] M. Ozuysal, V. Lepetit, F. Fleuret, and P. Fua, "Feature harvesting for tracking-by-detection.," in *ECCV (3)'06*, 2006, pp. 592–605.

[10] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Seventh IEEE Workshop on Applications of Computer Vision*, 2005, pp. 29–36.

[11] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vision*, vol. 77, pp. 125–141, May 2008.

[12] J. Wang, X. Chen, and W. Gao, "Online selecting discriminative tracking features using particle filter," in *Proc. CVPR*, 2005, pp. 1037–1042.

[13] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[14] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[15] G. Lathoud, J. M. Odobez, and D. Gatica-Perez, "Av16.3: an audio-visual corpus for speaker localization and tracking," in *Proceedings of the MLMI'04 Workshop*, 2004.