Recognition of people reoccurrences using bag-of-features representation and support vector machine

Kun Liu	extsuperscript{1}, Jie Yang	extsuperscript{2}

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China
E-mail: raistlin@sjtu.edu.cn	extsuperscript{1}, jieyang@sjtu.edu.cn	extsuperscript{2}

Abstract: In multi-camera surveillance systems, it is important to track the same person across multiple cameras. It is also desirable to recognize the individuals who have been previously observed in a single-camera system. The method that represents a object image using a bag of visual words has been commonly used in image retrieval applications. For recognizing people, it can outperform the methods mainly based on global appearance like color histogram, and fit better to low-quality images compared to biometric features such as face and gait. In this paper we study the details in feature extraction, vocabulary building and classifier learning of the bag-of-features approach for classifying tracks of different individuals. Based on this approach, we design a online system applying incremental support vector machine learning with a decision scheme to distinguish reoccurrences from new targets. We get promising results from the evaluation with more than 100 tracks of 50 different people.

Key Words: Tracking; multiple cameras; bag-of-features; support vector machine

1. INTRODUCTION AND RELATED WORK

In video surveillance system, to keep track of an individual is a base of event analysis and anomaly detection. Tracking people across multiple camera views can be divided into two parts, tracking in a single camera view and combining tracks of the same person by matching the targets in tracks. Then the behavior pattern can be better understood with the overall trajectories. Even for an isolated surveillance camera, the recognition of people reoccurrences is also useful, for it can detect suspicious lingering around that area and support tracking in recovery from failure.

While the methods based on overlapping view [1] and inter-camera space-time relationships [2] have their typical applications, they suffer some restrictions in more general scope. In contrast, feature matching is a wider applicable resort which is independent of overlapping region and path dependencies. As opposed to the biometric features like face or gait, we concentrate on general appearance features. Features based on color histogram are prevalent in recent work [2, 3]. Texture features extracted with specific filters like Gabor wavelets are also employed in object matching approaches [4], which use global statistics on descriptions of the image or match by alignment.

In contrast to these color or texture based methods which try to describe the global appearance of images, bag-of-features representations have recently become popular for image content based classification or retrieval owing to their simplicity and good performance. Local feature descriptors, which can achieve high robustness with respect to appearance variations, are employed to represent image content and all descriptors are quantized using learned visual words to facilitate the retrieval or classification [5, 6]. We believe that this description fits better to the high variation of object appearance across different views. Some local features can be very informative, therefore this approach can accommodate to bad localizations or part visibility.

The work by Hamdoun et al. [7] and Teixeira et al. [8] both use local feature descriptor to represent the target image. In Ref. [7], KD-tree built from visual words (local signatures) on interest points is used to serve the classification voting. However, KD-tree needs to be built with information from models of all tracked people, but the update and scalability problem is not addressed. Recent paper from Teixeira et al. [8] is an important reference for our final approach design. As proposed in their work, taking advantage of hierarchical k-means clustering to build a vocabulary tree can support large vocabulary. They introduced an SVM (support vector machine) based incremental learning algorithm to update the classifier for new class, but the evaluation result is not impressive. In this paper, we propose a online framework which use incremental SVM learning and two-level storage framework.

In Gray’s work [9] which uses Adaboost to find best viewpoint invariant representation from a group of simple features, it is shown that the color information has a large weight in the final selections. On the other hand, Teixeira et al. [8] demonstrate that their bag-of-features approach outperforms the major color histogram representation. In this context, we will try to improve discriminability by combining the color information into visual words.

From an application point of view, perhaps influenced by image retrieval applications, the most common scenario of previous local feature based work is using a given track or image to query previously similar ones [8, 9]. Their evaluations only focus on querying targets which have samples in storage. The problem here is that a full recognition approach should be able to find a test sample unmatchable to all classes in storage and label it as “unseen”. In other words, a good discrimination scheme does not lead to good recognition automatically if it
can not label a new class. For example, in Teixeira’s work [8] the supervised learning needs clear labels for new tracks but no decision scheme is presented, though it is implied that assistance from human operator is necessary. In this paper, we present a decision scheme to develop the high discriminability of the classifier to full recognition which can distinguish new targets from observed ones. This is important for some application scenarios like automatic detection of suspicious lingering.

This paper is structured as follows: Section 2 introduces the bag-of-features approach. Section 3 explains how we develop the discrimination among known classes to the full recognition. Section 4 describes our online system design and Section 5 presents experimental results. A final section gives the conclusion.

2. BAG-OF-FEATURES APPROACH

Bag-of-features approach first extracts patches from the people images and use local feature descriptors to describe them, then codes these descriptors by quantization against a priorily learned “visual word dictionary” (the vocabulary). This process tends to give the same label to similar local features. The occurrences of each visual word are counted to a histogram as a global representation of the image. We use support vector machines [10] to train a multiclass classifier on the histogram representations with their identity labels. Naturally, this classifier will label new images to the closest category. We develop its ability to estimate the true labels (which can be a new label) of tracks by combining similarity estimation. The system flowchart is shown in Fig.1.

2.1 Sampling Strategy

In the study of Nowak et al. [11], it is shown that random sampling gives equal or better classification than the sophisticated multi-scale interest operators that are in common use. We get the same notion from some simple test on people images. The interest region detectors like DoG (Difference of Gaussian) and MSER (Maximally Stable Extremal Regions) can only provide very few and unstable outputs in our dataset from compressed surveillance video. As the number of features extracted from each image is an important factor for classification performance, we believe random sampling is a better strategy here.

2.2 Structural Feature and Color Feature

The SIFT (Scale-invariant feature transform) [12] descriptor is employed to describe local features as it is widely used in image retrieval and classification. To combine color information, we try to add color description to tails of the original descriptor vectors. Of course, we need to deal with the disturbing illumination variance across views. A handy resource in video surveillance system is Cr and Cb channels in YCrCb color space. When we compute 16-bin histograms for each of the two chroma components and combine them into SIFT, we get extended 160-dimensional descriptor vectors. The evaluation result shows that SIFTCH and SIFT both have higher performance than color histogram and SIFTCH is better than SIFT especially in condition with large pose difference.

2.3 Vocabulary building

Considering high dimensions of descriptors and rich variance of people appearance, we believe a relatively large vocabulary is necessary. Using the hierarchical k-means clustering (HKM) to build a large vocabulary tree is proposed by Nistér et al. [13]. The tree construction is decided by two parameters, the number of centers at each level k and the number of levels I. It is built by initially producing k partitions with k-means clustering and then recursively cluster each partition to k sub-partitions until obtaining an l-level tree structure. Nodes are characterized by their partition centers. As proposed by Nistér et al. [13], we employ the Term Frequency Inverse Document Frequency (TF-IDF) scheme to weight the raw counts of visual words.

3. RECOGNITION USING SVM CLASSIFIER

For typical two-class classification in some feature space, support vector machines construct separating hyperplanes $\mathbf{w}\cdot\mathbf{x}+b=\pm 1$ in that space, which maximize the margin between the two data sets. As the number of features is large (according to the vocabulary size), using the linear kernel is enough and efficient. For the multiclass problem, it is usually implemented by reducing the original problem into multiple two-class sub-problems. The one-versus-all multiclass approach proves to be better in our evaluation. We implement linear SVM based on LIBLINEAR [14].

3.1 Develop discrimination to full recognition

As mentioned before, the multiclass SVMs and nearest neighbor classifiers in common use only fulfill the classification job in a limited range, that is, the test samples are assumed to belong to a known class. However, in our application, new classes will keep emerging. To distinguish a new class from the observed ones, we first use the trained classifier to get a ranking of learned classes according to the possibilities that the new instance belong to them, and then set a similarity threshold to select real similar classes as the
matching candidates. If none of the top rank classes get similarity higher than the threshold, we recognize the instance as a new individual which have never been observed. Otherwise, we label the instance to the best candidate class according to the obtained ranking. With this scheme, the full recognition considering new classes is obtained and the high discriminability can also be achieved by supervised learning.

It is efficient and simple to use the same measurement for similarity threshold and classifier training. Direct distance measurement against class centers proves to be able to serve as a classifier for BoF vectors (Section 5.2) and the distances is a direct reflection of similarity. So this is a simple baseline method.

3.2 Full recognition scheme with SVM

Now we develop the similarity threshold scheme for SVM. When training the submodel for a class \( c_n \) in one-versus-all multiclass SVM, we set the labels of samples in \( c_n \) as “+1”. Therefore the samples in \( c_n \) lie in the positive side of the hyperplane \( \mathbf{w}_n \cdot \mathbf{x} + b_n = 0 \), and samples from other classes in training set \( (c_{other}) \) should lie in the negative side. A winner-takes-all voting for classification means assigning the sample \( \mathbf{x} \) to the class with maximum value of \( \mathbf{w}_n \cdot \mathbf{x} + b_n \). In fact, this can be seen as a nearest neighbor classifier using \( f(\mathbf{x},c_n) = \mathbf{w}_n \cdot \mathbf{x} + b_n \) as a similarity measure, which tells how the sample \( \mathbf{x} \) conforms to the learned boundary between class \( c_n \) and other classes. Our full recognition scheme also considers \( \mathbf{w}_n \cdot \mathbf{x} + b_n \) as the similarity between a sample and the class \( c_n \). And we make an approximation by taking \( c_{other} \) in a sub-problem as groups of samples for \( c_n \). Then for the samples from unseen classes, we also expect them to likely lie in the negative side of the learned boundary. The formula for the full recognition scheme is:

\[
\begin{align*}
\text{probability}(\mathbf{x} \in c_n) &\propto f(\mathbf{x}, c_n) \\
\max_n(f(\mathbf{x}, c_n)) &< T_s \Rightarrow \mathbf{x} \in c_{new} \\
\end{align*}
\]

The best \( T_s \) will be selected by experiment. When processing a track that contains multiple samples, we take the average of \( f(\mathbf{x}, c_n) \) calculated for all its samples as \( \text{sim}(c_n) \), the similarity of the track to \( c_n \). Then we can get a ranking according to \( \text{sim}(c) \) for all classes. By inserting an item \( c_{new} \) with \( \text{sim}(c_{new}) = T_s \), we obtain a ranking with a new-class label involved. When \( c_{new} \) is the top ranking class, the track should be assigned a new label.

4. ONLINE SYSTEM DESIGN

A online system requires that we can deal with the new information coming with captured tracks, including new samples for both known classes and unknown classes. Teixeira et al. [8] make use of an ensemble based incremental algorithm with SVM as the base classifier. However, its performance is far lower than the model built with complete data. Considering that the model itself needs to store a high dimension hyperplane norm vector while the BoF vectors are very sparse with only hundreds of nonzero elements, to store part of previous data is not a very consuming approach in comparison. Based on this fact, we implement the incremental learning by saving support vectors [15]. To restrict the amount of data stored in memory, we may only store the support vectors with large contributing weight for the norm vector \( \mathbf{w} \) (This weight can be measured by Lagrange multipliers \( \alpha_i \) in the Lagrangian formulation of SVM [10]).

We extend the Error-driven technique to incrementally update the learned multiclass SVM classifier. The Error-driven technique means that we keep misclassified samples and use them together with the stored support vectors to obtain a new classifier. In the one-versus-all multiclass SVM, instead of rebuilding all submodels, only the submodels which make wrong decisions for the new samples are updated with the misclassified samples and stored support vectors. \( \mathbf{x} \) is considered as misclassified by the submodel of \( c_{new} \), if \( f(\mathbf{x}, c_{new}) > \min(T_s, 0) \) for \( \mathbf{x} \in c_{new} \), or \( f(\mathbf{x}, c_{new}) < \max(T_s, 0) \) for \( \mathbf{x} \in c_{new} \). For a new class, we train a new submodel and update other submodels which have misclassification on the new samples. Through this error-driven incremental SVM learning, we resolve the out-voting problem by updating the disturbing submodels with the disputed samples.

As the new classes keep emerging, the scale of the multiclass model expands all along. Therefore the computation burden will be unacceptable for learning and classification sooner or later. So a constraint needs to be enforced on the number of submodels. By recording the time stamps of tracks, we can set the maximum duration before discarding a class since the last time it has been observed, or set the maximum allowed number of submodels. When this limit is violated, we discard the class which has not been observed for the longest time and collect the information of these classes (submodel and support vectors) into some data files when they are eliminated off memory. This two-level storage framework allows us to do wide search in system idle time by accessing those data files and take this extracted information as video indexes. With this framework as complement, our design is a totally scalable system.

5. EXPERIMENT

5.1 Dataset

We employ CAVIAR dataset for evaluations. The dataset contains video sequences in corridors of a commercial mall (coming from the EC Funded CAVIAR project/IST 2001 37540, found at URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/). We extract several discontinuous short tracks from the long sequences, and they can be treated as different occurrences for they are just different in scale, background and illumination conditions. We treat each short track as a unit and evaluate the recognition by separating them into training set and test set. Random selected 47 pairs of tracks from the dataset are represented by DC-47. Another dataset VIPeR which contains pedestrian image pairs [16] is only employed to build vocabulary. In these evaluations, we will use the very coarse segmentation result, the bounding box of objects.
5.2 Evaluations for bag-of-features approach

To evaluate different selections of parameters in the bag-of-features approach, we test the recognition rates for tracks (based on voting) by averaging on multiple times of 2-fold cross validations with DC-47. We list some test results in Table 1. Since we need to use a priorly learned vocabulary in the actual operation, we also test the effect of quantization against the vocabulary trained from VIPeR dataset. Based on these results, we will use the SIFTCH descriptor and a vocabulary tree (k=20, l=3) trained from VIPeR dataset in following experiments.

Though the color information is less discriminative, it is more robust with respect to viewpoint change than local structural features because previously captured features may become invisible in current viewpoint. We find that the benefit to combine color information is more evident for the pairs of tracks which contain large viewpoint changes. Usually denser sampling will promote the recognition rate at the cost of efficiency. Based on experiments, we believe that extracting 20~30 samples for each track and at least more than 200 patches from each sample image is enough for precision.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>k</th>
<th>level</th>
<th>TR (%)</th>
<th>TR* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>10</td>
<td>10000</td>
<td>90.9</td>
<td>87.4</td>
</tr>
<tr>
<td>SIFT</td>
<td>20</td>
<td>8000</td>
<td>90.3</td>
<td>88.7</td>
</tr>
<tr>
<td>SIFTCH</td>
<td>20</td>
<td>8000</td>
<td>93.0</td>
<td>92.1</td>
</tr>
<tr>
<td>SIFTCH</td>
<td>10</td>
<td>10000</td>
<td>91.8</td>
<td>87.4</td>
</tr>
<tr>
<td>SIFTCH</td>
<td>10</td>
<td>1000</td>
<td>89.6</td>
<td>89.5</td>
</tr>
<tr>
<td>SIFTCH</td>
<td>5</td>
<td>125</td>
<td>82.1</td>
<td>79.1</td>
</tr>
</tbody>
</table>

*quantized against the vocabulary trained from VIPeR dataset

Table 2: Track recognition rate for different approaches

<table>
<thead>
<tr>
<th>Representation</th>
<th>Classifier</th>
<th>TR (%)</th>
<th>1D CH</th>
<th>HLH</th>
<th>HLH</th>
<th>BoF</th>
<th>BoF</th>
<th>BoF</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR (%)</td>
<td>KNN</td>
<td>57.2</td>
<td>60.1</td>
<td>72.3</td>
<td>86.5</td>
<td>90.5</td>
<td>93.0</td>
<td></td>
</tr>
</tbody>
</table>

To demonstrate the advantage of our approach, we test simple 1D color histogram (1D CH) and hand localized 1D-ColorHistogram (HLH) [16] with KNN (get best performance using Bhattacharyya distance and 128 bins for each channel) or SVM classification. It is shown that SVM can help select out dominant features. On the other hand, classification using KNN or class centers based on BoF representation (Bhattacharyya distance provides better performance than 1-norm distance) also provide good performance, which reveals that our approach mainly depends on the advantage of BoF. Relevant evaluation results list in Table 2. To summarize, the BoF approach is very effective representation for recognizing people across different view and organize test sets with both reoccurring people and new ones by multiple times of random selections. By checking the distributions of similarity measurements for these test sets, we can choose a suitable threshold. Fig.2 shows some typical distributions of SVM weighting function \( f(x,c_n)=w_n^T x + b_n \) in three conditions with different number of training classes. The results support our idea of taking \( c_{\text{others}} \) as a approximation of \( c_n \), since the distribution of \( f(x,c_n) \) for \( x \) in \( c_{\text{others}} \) and the distribution for \( x \) in unobserved classes \( (c_n \setminus c_{\text{others}}) \) are close. For the concern about the universality of a constant threshold, we find that the distributions are relatively stable when we have enough classes to serve as \( c_{\text{others}} \) for submodel training. This requirement can be met by an initialization stage. To evaluate the effectiveness of the threshold as a standard to recognize reoccurrence and new ones, we check the recall/fall-out rate for detection of never-seen individuals in test sets (Table 3). The result suggests “pushing” the hyperplane toward \( c_n \) to give \( c_n \) larger margin for better trade-off. Accordingly, we choose \( T_s=0.3 \) for our system test. When we check the distribution of Bhattacharyya distance between track samples and class centers, we get a similar result and \( T_s=0.8 \) is selected.

5.3 Full recognition and System evaluation

To test our full recognition scheme, we first use random selected tracks to construct a multiclass model, and then

<table>
<thead>
<tr>
<th>Number of classes = 30</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>f(x,c)</td>
<td>0.05</td>
</tr>
<tr>
<td>Learned others</td>
<td>0.1</td>
</tr>
<tr>
<td>Unobserved</td>
<td>0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of classes = 45</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>f(x,c)</td>
<td>0.05</td>
</tr>
<tr>
<td>Learned others</td>
<td>0.1</td>
</tr>
<tr>
<td>Unobserved</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Figure 2 Histogram of \( f(x,c_n) \) in three conditions

To evaluate the system performance for long-time running with the effect of incremental learning and scale control, we put all 150 short tracks of 59 different individuals together to simulate input track sequences. In this evaluation, a maximum of 30 support vectors are reserved for one class in incremental learning and a maximum of 40 submodels are allowed in the multiclass SVM. For similar targets and complex conditions, assistance from human operator is necessary for high matching accuracy. Meanwhile, automation is also desirable in surveillance system. So we test the system for two operation manners (Fig.3). The probability that the correct
match (includes assigning new-class labels for unseen individuals) has a rank equal to or less than some value is plotted. Some typical ranking results are shown in Fig 4. From these curves, we find that the SVM based approach with manual decisions give best performance as expected. It is also notable that the SVM approach suffers more from wrong classification decisions. The actual operating situation may be some kind of combination of these two operation manners. Systems using our approach can keep running without supervision and the operator’s assistance can improve the recognition accuracy.

Table 3: Recall/fall-out of detection of never-seen individuals for different thresholds

<table>
<thead>
<tr>
<th>$T_s$</th>
<th>15 classes</th>
<th>30 classes</th>
<th>45 classes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>recall</td>
<td>fall-out</td>
<td>recall</td>
</tr>
<tr>
<td>-0.4</td>
<td>0.40</td>
<td>0.08</td>
<td>0.65</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.63</td>
<td>0.13</td>
<td>0.78</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.73</td>
<td>0.21</td>
<td>0.88</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.84</td>
<td>0.25</td>
<td>0.92</td>
</tr>
<tr>
<td>0</td>
<td>0.91</td>
<td>0.36</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Fig. 3: CMC curves for the two recognition approaches

Fig. 4: Typical ranking results (for the targets in the left) One sample is shown for each member tracks of the top ranking classes

6. CONCLUSION

In this paper, we have proposed an approach for recognizing people reoccurrence in view-independent tracks. Based on the bag-of-features representation, a scalable online recognition framework is presented. We combine the supervised learning with similarity measurement to build a full recognition scheme, which can distinguish new classes from observed ones. Our approach seems promising in evaluation and may be beneficial in other areas like general object recognition and feature based object tracking.

ACKNOWLEDGEMENTS

This study is supported by the National Natural Science Foundation of China (60675023).

REFERENCES