Structural Local DCT Sparse Appearance Model for Visual Tracking

B. K. Shreyamsha Kumar, M. N. S. Swamy and M. Omair Ahmad

© IEEE 2015. All rights reserved.

This paper is published in proceedings of IEEE International Symposium on Circuits and Systems (ISCAS) and the original publication is available at http://ieeexplore.ieee.org (DOI: 10.1109/ISCAS.2015.7168853). 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. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for commercial purposes, or modification of the content of the paper are prohibited and require prior specific permission and/or a fee.

Cite this article as:

Bibtex:
@Inproceedings{SLDAMVT2015,
    Title = {Structural Local \{DCT\} Sparse Appearance Model for Visual Tracking},
    Author = {Shreyamsha Kumar, B. K. and Swamy, M. N. S. and Omair Ahmad, M.},
    Booktitle = {Proc. of the IEEE Int. Symp. on Circuits and Systems (ISCAS)},
    Year = {2015},
    Month = {May},
    Pages = {1194--1197},
    DOI = {10.1109/ISCAS.2015.7168853},
}
Structural Local DCT Sparse Appearance Model for Visual Tracking

B. K. Shreyamsha Kumar, Student Member, IEEE, M. N. S. Swamy, Fellow, IEEE, M. Omair Ahmad, Fellow, IEEE
Department of Electrical and Computer Engineering
Concordia University
Montreal, Quebec, Canada
{s_bidare, swamy, omair}@ece.concordia.ca

Abstract—The success of sparse representation in face recognition has motivated the development of sparse representation-based appearance models for visual tracking. These sparse representation-based trackers show state-of-the-art performance, but at the cost of computationally expensive $l_1$-norm minimization. As the computational cost prevents the tracker from being used in real-time systems such as real-time surveillance and military operations, it has become a very important issue. With the aim of reducing the computational complexity of $l_1$-norm minimization, a structural local DCT sparse appearance model is proposed in a particle filter framework. Application of DCT on local patches helps to reduce the dimensions of the dictionary as well as candidate samples by using low-pass filtered DCT coefficients. This in turn helps to remove the information relating to occlusion and background clutter thereby reducing the ambiguity created while computing the confidences of the target samples. The proposed method is evaluated on the challenging image sequences available in the literature and its performance compared with three recent state-of-the-art methods. It is shown that the proposed method provides superior/similar performance for most of the sequences with reduced computational complexity in $l_1$-norm minimization.

Keywords—visual tracking; appearance model; local DCT; $l_1$-norm minimization; template update

I. INTRODUCTION

In the last two decades, visual object tracking has seen a flurry of research activities due to its wide range of real-life applications including vehicle navigation, robotics, human behavior analysis, action recognition, human computer interaction, video indexing and retrieval, medical imaging, security and surveillance [1, 2]. In general, the task of tracking is formulated as a searching problem that finds a most probable candidate having a similar appearance as that of the given object [3]. It still remains a challenging problem due to its complexity in target searching as well as intrinsic and extrinsic object appearance variations [4]. In order to achieve a reliable tracking performance, a good appearance model that adapts to intrinsic appearance variations and is robust to extrinsic appearance variations, is required.

In the literature, the tracking algorithms are categorized into either generative or discriminative approaches based on the object representation scheme. Discriminative methods extract information from both the target and the background to differentiate the target from the background and hence, it is formulated as a binary classification problem [5-12]. Grabner et al. [5] proposed an online boosting algorithm for tracking by updating the discriminative features and later proposed a semi-online boosting algorithm, which treats all the tracking results as unlabeled data and adapts a classifier within the semi-supervised framework [6]. Subsequently, Babenko et al. [7] proposed a multiple instance learning (MIL) algorithm using all the ambiguous positive and negative samples in the bags to train a discriminative model to reduce the drifting problem. Kalal et al. [8] exploited the underlying structure in positive and negative samples to learn a P-N classifier for object tracking. In [9], Wang et al. proposed a discriminative tracking algorithm using superpixels to differentiate the target from the background. Later, a two-stage online discriminative algorithm was proposed using local sparse codes to represent the target object and exploiting both the static and adaptive observation models [10]. Zhang et al. [11] proposed an online discriminative feature selection method by directly coupling the classifier score with the sample importance, and optimizing the objective function in the steepest ascent and descent direction for positive and negative samples, respectively. Also, a compact 3D-DCT-based object representation scheme and its incremental learning for robust visual tracking have been proposed in [12] using a signal reconstruction-based similarity measure to evaluate the likelihood.

Generative methods extract information only from the target region to model the target appearance and search for a region that is most similar to the target model. Wang et al. [13] proposed an adaptive appearance model in a joint spatial-color space using Gaussian mixtures. Ross et al. [14] proposed a linear subspace learning to model the target appearance with a sample mean update to track the object in a particle filter framework. In [15], wide variations of illumination and pose are captured by decomposing the observation model into multiple basic observation models. Mei et al. [16] proposed a sparse representation of the target and trivial templates as the appearance model to represent the object and occlusion information respectively, and the object is tracked by solving $l_1$-norm minimization problem. In order to speed up the tracking process, the less expensive $l_2$-norm minimization has been exploited [17] to bound the $l_1$ approximation error without sacrificing accuracy. Liu et al. [18] proposed a tracking algorithm based on local sparse model using histograms of
sparse coefficients and mean-shift algorithm. Jia et al. [19] proposed an adaptive structural local sparse appearance (ASLA) model for tracking by exploiting both partial and spatial information of the target based on an alignment-pooling method.

For most of the high-level applications, such as action recognition and retrieval, real-time tracking speed is a practical requirement and hence, the computational complexity has become a very important issue. The sparse representation-based appearance models show good performance, however, at the computational expense of $l_1$-norm minimization. In order to reduce the computational complexity of $l_1$-norm minimization, in this paper, a structural local DCT sparse appearance model is proposed. The proposed method extracts overlapped local image patches within the target region and computes the 2D-DCT of all the patches followed by zigzag scanning to create a local DCT coefficients vector. The energy compaction property of the DCT is exploited to reduce the dimension of the dictionary as well as that of the candidate samples. These reduced dictionary and candidate samples are used to obtain sparse codes of the local patches, and the alignment-pooling of the sparse codes is used as the similarity measure to find the target location. The paper is organized as follows. Section II describes the proposed method, and Section III presents the experimental results. Section IV summarizes the work of this paper and gives briefly a plan for the future work.

II. Proposed Algorithm

In this section, motivated by the success of local sparse representation for object appearance [10, 18-20], a generative model of object representation scheme using sparse codes of local DCT patches is proposed to reduce the computational cost of $l_1$-norm minimization by exploiting the energy compaction property of DCT. The proposed algorithm has some similarity to ASLA [19] in the use of local sparse representation and template update scheme, but differs in the domain in which sparse representation is applied. ASLA directly uses pixels in the local patches for sparse representation but the proposed method uses the DCT coefficients of the pixels in the local patches for sparse representation. Use of DCT serve two purposes, 1) dimensionality reduction of the dictionary and candidate samples by low-pass filtering the DCT coefficients and hence, reduction in computations, 2) low-pass filtering helps to remove the details pertaining to occlusion and background clutter, which create ambiguity while computing the confidence of the target samples.

The block diagram of the proposed method in a particle filter framework is shown in Fig. 1. Here, motion of the object is estimated using a Markov model with hidden state variables [21]. Let $Z_t$ represent a state variable describing the affine motion parameters (the location and pose) of a target at time $t$. Given a set of observed images $O_t = \{O_1, ..., O_t\}$ at time $t$, the posterior probability $p(Z_t|O_t)$ can be inferred recursively by the Bayesian theorem,

$$p(Z_t|O_t) \propto p(O_t|Z_t) \int p(Z_t|Z_{t-1})p(Z_{t-1}|O_{t-1})dZ_{t-1}$$  \hspace{1cm} (1)

where $p(O_t|Z_t)$ represents the observation model, and $p(Z_t|Z_{t-1})$ represents the dynamic (state transition) model. In the tracking framework, the dynamic model $p(Z_t|Z_{t-1})$ describes the temporal correlation of states between two consecutive frames and is modeled by Gaussian distribution, i.e., $p(Z_t|Z_{t-1}) = \mathcal{N}(Z_t; Z_{t-1}, \Sigma)$, where $\Sigma$ denotes a diagonal covariance matrix whose elements are the variances of the affine parameters. The dynamic model in the motion estimation block of Fig. 1 randomly selects $M$ samples of the state variable $Z_t$ at time $t$ given the state $Z_{t-1}$ at time $t-1$. From these $M$ states, $Z^m_t$, $M$ target candidates $O^m_t$ are generated, where $m = 1, 2, ..., M$.

For a given target candidate, the overlapped local patches inside the target region are extracted with a spatial layout and 2D-DCT of these patches are computed. The DCT coefficients of each patch are low-pass filtered after zigzag scanning to reduce its dimension from $d$ to $r$, where $d$ is the dimension of the local patch vector before filtering and $r$ is the dimension of the local patch vector after filtering. This gives a vector $Y = \{y_1, y_2, ..., y_N\} \in \mathbb{R}^{r \times N}$ for a given candidate after low-pass filtering. Similar procedure is followed for every template in a given set of target templates $T = [T_1, T_2, ..., T_N]$ to create a dictionary $D = \{d_1, d_2, ..., d_{(N \times N)}\} \in \mathbb{R}^{r \times (N \times N)}$, where $N$ is the number of local image patches extracted within the target region and $n$ is the number of target templates. Each fixed part of the target object is represented by one local patch and hence, all these local patches with a fixed spatial relationship can completely represent the target structure. Further, as these local patches are obtained across many templates, the resultant dictionary captures the generality of the different templates and hence, it is able to represent various forms of target parts [19].

In sparse representation, only a few basis elements of the dictionary with different coefficients are sufficient to represent the local patches inside the target region and this is achieved by solving the following minimization problem:

$$\min_{b_i} \|y_t - Db_i\|_2^2 + \lambda \|b_i\|_1, \text{ s.t. } b_i \geq 0, \quad (2)$$

where $y_t \in \mathbb{R}^{r \times 1}$ represents the low-pass filtered 2D-DCT of the $i$-th local patch vector, $b_i \in \mathbb{R}^{(N \times N) \times 1}$ is the corresponding sparse code of that local patch, and $b_i \geq 0$ indicates that all the elements of $b_i$ are non-negative. Now, the sparse codes of the given target candidate $Y$ is given by $B = [b_1, b_2, ..., b_N]$, where the sparse coefficients of each local patch $b_i$ are divided into several segments depending on the template that each element of the vector belongs to, i.e., $b_i^T = [b_{i1}^{(1)}, b_{i2}^{(2)}, ..., b_{in}^{(n)}]^T$, where $b_i^{(k)} \in \mathbb{R}^{n \times 1}$ indicates the $k$-th segment of the sparse coefficient vector $b_i$ corresponding to the template $T_k$ in the

Fig. 1. Block diagram of the proposed method (refer e-version for color)
given target template set $T$. From these segmented sparse coefficients $b^{(k)}$, a normalized feature vector $v_i = \frac{1}{C} \sum_{k=1}^n b_i^{(k)}$, $i = 1, 2, ..., N$, where $C$ is a normalization term. With this strategy, all the normalized feature vectors of the local patches within a candidate region form a square matrix $V = [v_1, v_2, ..., v_N] \in \mathbb{R}^{N \times N}$. Note that each local patch at a certain position of the candidate is represented by patches at different positions of the templates. But, the local appearance of a patch in a candidate is correctly represented by the patches at the same positions of the template. This is taken care by considering only the diagonal elements of the square matrix $V$ as the pooled feature vector $f = \text{diag}(V) \in \mathbb{R}^{N \times 1}$. This feature vector captures the target structure with a fixed spatial relationship and reflects the similarity between the candidate and the target template. Hence, the observation model $p(O^m | Z^m)$, the likelihood of the observation $O_t^m$ at state $Z_t^m$, is represented by the pooled feature $f^m$ as $p(O_t^m | Z_t^m) \propto \sum_{i=1}^N f_i^m$.

In the decision making block, the optimal state of the tracked target $Z_t^m$ is determined by solving the maximum posteriori estimation problem:

$$\hat{Z}_t = \arg\max_{Z_t^m} p(O_t^m | Z_t^m) p(Z_t^m | Z_{t-1})$$

Finally, the template set is updated to take care of the object appearance variations using incremental subspace learning and sparse representation, as done in ASLA [19]. Once the template is updated, the low-pass filtered DCT coefficients of the overlapped patches pertaining to the updated template are used to update the dictionary. The template and dictionary update reduces the drifting problem as well as the influence of the template with partial occlusion.

### A. Computational Complexity

Computational complexity is a very important issue as a high complexity prevents the tracker from being used in real-time applications such as real-time surveillance and military operations. Since $l_1$-norm minimization is a most computationally expensive operation for sparsity-based trackers, both the number of $l_1$-norm minimizations and the cost of solving each $l_1$-norm minimization significantly affect the computational speed of the trackers. In order to reduce the computational complexity, the proposed method aims at reducing the cost of each $l_1$-norm minimization by exploiting the energy compaction property of DCT. As the proposed method is similar to ASLA except for the use of DCT on local patches, the computational complexity and run-time (for woman sequence) of $l_1$-norm minimization in the proposed method is compared with that of ASLA in Table I. From Table I, it is observed that $l_1$-norm minimization in the proposed method is faster than that of ASLA by a factor of $\frac{d}{c}$. Also, the run-time of $l_1$-norm minimization in the proposed method including DCT computation is less than that of ASLA. Note that, the run-time does not reflect the actual computations as it depends on number factors such as way of coding (non-optimized code), applications running in the background on the computer etc.

### III. Experimental Results

The proposed algorithm is implemented in Matlab and its performance is evaluated using twelve challenging sequences [4, 7, 14, 15, 19, 20]. These sequences cover most of the real-life challenging situations in object tracking such as varying lighting conditions, low contrast, motion blur due to fast movement, pose variation, complex background, heavy occlusion, scale change, and in-plane and out-of-plane rotation. To compare the performance of the proposed method, recent state-of-the-art algorithms, ASLA (generative model) [19], SPT (generative model) [4] and SCM (collaborative model) [20] are considered. For a fair comparison, the source codes provided by the authors of these methods are used by keeping their initialization and other parameters intact. For all the trackers, the location of the target object is labeled manually in the first frame. In the proposed method, each image observation is resized to 32 × 32 pixels and then local patches of size 16 × 16 are extracted with an overlap of 8 pixels. For all the sequences, the dimension of the local patch vector after filtering, $r$, is set to 64. In the proposed method, SPAMS package [22] is used for $l_1$-norm minimization and the regularization constant $\lambda$ is set to 0.01. Considering the trade-off between effectiveness in tracking and computational efficiency, 600 particles are sampled using a particle filter.

For quantitative performance comparison, two popular evaluation criteria namely, center location error (CLE) and overlap rate (OR) [1] are employed. For a better tracking performance the value of CLE should be close to 0, and OR should be close to unity. Table II summarizes the tracking results of the twelve sequences in terms of average OR and CLE with two best results shown in red and blue colors indicating the first and second best performances, respectively. It is observed that the performance of the proposed tracker is better in some sequences (Animal, Car11, Cliffbar, Girl, Jumping and Woman) and comparable in other sequences (Car4, DavidIndoor, Faceocc2, Occlusion1, Signer1, Stone). This performance is achieved with reduced computations in $l_1$-norm minimization due to filtering DCT coefficients. From Table II, it can be observed that sometimes the tracker standing first in average OR may not be first in terms of the average CLE. For example, in Animal, Jumping and the Woman sequences, the tracker standing first in terms of the average CLE came second in terms of average OR and vice-versa. This creates ambiguity while deciding the tracking performance. In contrast to CLE that captures only the relative positions between the centers, OR that captures the scale variations of the object through the bounding boxes of the tracking result along with positions of the object, enjoys superiority in evaluating the tracking performance. Note that the proposed method performs well and is competitive for most of the sequences in terms of average OR and CLE. Fig. 2 shows the effect of filtering DCT coefficients on OR and it can be observed that OR is higher for most of the frames with 64

![Fig. 2. Effect of filtering DCT coefficients on OR (refer e-version for color)](image-url)
DCT coefficients as compared to that with 128 and 256 DCT coefficients. Due to lack of space the qualitative comparison is not given here; readers are requested to visit the site http://youtu.be/O9s5xVEp1Tc.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, a structural local DCT sparse appearance model has been proposed in a particle filter framework to reduce the dimension of the dictionary as well as that of candidate samples by exploiting the energy compaction property of DCT. Low-pass filtered local DCT coefficients have not only reduced the computational complexity of $l_1$-norm minimization by a factor of $d/r$, but also have aided in removing the information relating to occlusion and background clutter thereby reducing the ambiguity created while computing the confidences of the target samples. The performance of the proposed method has been compared with that of three recent state-of-the-art methods and it has been shown that the proposed method yields superior/similar performance for most of the sequences with reduced computations.

In the proposed method, the dimension of local DCT vector after filtering is set to 64 for all the patches irrespective of the information in the patch. The adaptability of this dimension as per the information in the respective patch to exploit the energy compaction property of DCT is yet to be explored. Also, the application of DCT in template update scheme to reduce the computations of the $l_1$-norm minimization needs to be investigated.

<table>
<thead>
<tr>
<th>Methods</th>
<th>ASLA</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$l_1$-norm</td>
<td>$l_2$-norm</td>
</tr>
<tr>
<td></td>
<td>minimization</td>
<td>minimization</td>
</tr>
<tr>
<td>One sample</td>
<td>$O(dnN^2)$</td>
<td>$O(rnN^2)$</td>
</tr>
<tr>
<td>Woman sequence</td>
<td>270.84</td>
<td>221.05</td>
</tr>
</tbody>
</table>

TABLE II AVERAGE OR AND CLE

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Average OR</th>
<th>Average CLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASLA</td>
<td>SPT</td>
</tr>
<tr>
<td>Animal</td>
<td>0.615</td>
<td>0.607</td>
</tr>
<tr>
<td>Car4</td>
<td>0.905</td>
<td>0.922</td>
</tr>
<tr>
<td>Car11</td>
<td>0.825</td>
<td>0.808</td>
</tr>
<tr>
<td>Cliffbar</td>
<td>0.457</td>
<td>0.543</td>
</tr>
<tr>
<td>DavidIndoor</td>
<td>0.757</td>
<td>0.800</td>
</tr>
<tr>
<td>Faceocc2</td>
<td>0.822</td>
<td>0.836</td>
</tr>
<tr>
<td>Girl</td>
<td>0.716</td>
<td>0.593</td>
</tr>
<tr>
<td>Jumping</td>
<td>0.683</td>
<td>0.687</td>
</tr>
<tr>
<td>Occlusion1</td>
<td>0.879</td>
<td>0.913</td>
</tr>
<tr>
<td>Singer1</td>
<td>0.797</td>
<td>0.822</td>
</tr>
<tr>
<td>Stone</td>
<td>0.506</td>
<td>0.663</td>
</tr>
<tr>
<td>Woman</td>
<td>0.751</td>
<td>0.101</td>
</tr>
</tbody>
</table>

REFERENCES


