Robust visual tracking with ‘structured sparse’ representation appearance model [1]
Challenge: tracking the non-stationary appearance of objects
- significant pose, illumination variations and occlusions

A visual tracking algorithm can be decomposed into three components:

- An appearance model
  - Captures the visual characteristics of the target and evaluates the similarity between observed samples and the model;

- A motion model
  - Locates the target between successive frames with certain motion hypothesis;

- An optimization strategy
  - Associates the appearance model with the motion model and finds the most likely location in the current frame

In this work, focus is on a robust appearance model
- That confronts the aforementioned difficulties.
Static subspace representation aims at representing the target with a low-dimensional subspace based on the core desire for dimensionality reduction.

However, these methods mainly rely on training before tracking that may lead to failure if the target experiences variations which are outside the set of training samples.

Moreover, in many real applications, it is neither practical to construct a rich sample set nor perform extensive training off-line.

What is needed, therefore, is an adaptive model that has the capabilities of online learning and updating for time-varying appearance representations.

Ross et al. propose an incremental Principal Component Analysis (PCA) based learning subspace model that shows robustness to gradual changes in pose, scale and illumination.
However, the above algorithms, static or adaptive, usually do not have mechanisms to handle occlusions and they could suffer from failure caused by occlusion in a long duration.

Mei and Ling [2] propose a sparse representation based appearance model in a visual tracking scenario,

- The target appearance is expressed as a sparse linear combination with a basis library consisting of target templates, trivial templates via $l_1$ minimization.

The tracking result is then assigned to the observed sample that has the smallest reconstruction residual with the target templates and corresponding target coefficients.

Qualitative experiments exhibit impressive robustness of such approach, but the large computational load prohibits its further application in reality.
Related work and drawbacks - Summary

- Traditional appearance models - template or subspace representations
  - Focus on approximating the target appearance itself,
  - Sensitive to gross errors caused by occlusion.

- Recent, Sparse representation based appearance modeling
  - Attempt to jointly estimate the target appearance as well as the occlusion
  - Shown to give robustness against various disturbances, particularly, in the sense of occlusion
  - The $l1$ tracker, has extensive computational cost that the tracker is not readily applicable to practical implementations.
What is proposed new

- Structured sparse representation appearance model
  - The model goes beyond simple sparsity by considering a priori information on the predefined structure of the basis library and contiguous spatial distribution of occlusions.

- In this model, the nonzero entries in the sparse coefficient vector have a particular structure that arises from practical visual tracking.

- The appearance model is cast as a sparse linear combination of a structured union of subspaces instead of individual templates.

- It is shown that with this structured model, a more robust and efficient implementation of the sparse representation for visual tracking is feasible.

- Introduce the Block Orthogonal Matching Pursuit (BOMP) rather than $l_1$ minimization or Orthogonal Matching Pursuit (OMP)

- Exploit the inherent discriminativity of BOMP to eliminate the invalid observations during tracking, which results in more accurate and efficient tracking results
Sparse representation based appearance model

- Consider the observed target sample \( y \in \mathbb{R}^L \)

- Approximately lies in a subspace spanned by \( d \) given training templates

\[
y \approx A_T x = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d,
\]

Training templates \( A_T = [a_1 a_2 \ldots a_d] \in \mathbb{R}^{L \times d} (L > d) \)

Target Coefficient vector \( x = (x_1, x_2, \ldots, x_d)^T \in \mathbb{R}^d \)

- With unpredictable partial occlusion

\[
y = A_T x + e = [A_T A_e][\begin{pmatrix} x \\ e \end{pmatrix}] = A \omega.
\]

- \( \omega \) is a sparse coefficient vector with non-uniform density, where the \( e \) part is sparse, but the \( x \) part is usually dense
$l_0, l_1$ minimization and greedy algorithms

- $x$ and $e$ can be jointly recovered by solving the $l_0$ norm minimization
  \[
  \min_{\omega} \| \omega \|_0 \quad \text{subject to} \quad \| y - A\omega \|_2 < \epsilon,
  \]
  where $\| \cdot \|_0$ is the $l_0$ norm, $\| \cdot \|_2$ is the $l_2$ norm, and $\epsilon$ is the level of reconstruction error.

- But this problem is numerically unstable and NP-hard.

- The $l_1$ tracker converts this problem into an equivalent $l_1$ norm convex optimization problem.

  - Solvable in polynomial time, but still complex and costly.

- Alternate methods – Matching Pursuit (MP) and Orthogonal Matching Pursuit (OMP).

  - Superior in efficiency and implementation.
  - Comparable results to $l_0$ minimization.
- The sparse supports of $\omega$ often have an underlying order
- Partial occlusion often appears as a contiguous spatial distribution in the observed target sample
- It was shown that analyzing the local image patches can promote the recognition performance against partial occlusion
Local region analysis & Block sparsity

- A local region based method to resolve the partial occlusion problem

- First partition the observed sample and each of the training templates into local parts

- The contiguous occlusion can be encoded as a block sparse vector having clustered nonzero entries

- Recent works report impressive improvements of the recovery threshold and efficiency by considering the block structured sparsity
A block $k$-sparse vector $\omega$ is defined as the vector that has at most $k$ blocks with nonzero Euclidean norm.
PCA and Eigen space

- Orthogonalizing the columns within the blocks is shown to achieve a higher recovery threshold.

- The occlusion template set is the identity matrix and so satisfies this condition as all the columns in the identity matrix are orthogonal mutually.

- The target template set, however, is usually not orthogonal, but has a high coherence because the templates are similar to each other.

- Thus, we exploit the Principal Components Analysis (PCA) approach to orthogonalize the target template set and form an Eigen template set instead.

\[ y = [UA_e][ \begin{array}{c} c \\ e \end{array} ] = D\alpha, \]

- The Eigen template set \( \hat{U} \) is obtained by Singular Value Decomposition (SVD)

\[ A_T = U\Sigma V^T \]
Block Orthogonal Matching Pursuit (BOMP)

- When the basis library $A$ and sparse coefficient vector $x$ are block structured, a more efficient implementation of MP and OMP is possible – Block OMP

- The BOMP algorithm consists of three major stages in each iteration:
  - matching stage, estimation stage and updating stage.

- The major difference between BOMP and standard OMP is the matching stage.
  - BOMP selects the block having the highest correlation whereas only one best matched template is chosen by OMP.

- Once the block is found, the corresponding coefficients are estimated via least-squares minimization at the estimation stage. Then the residual is updated.
Algorithm: BOMP with outlier elimination

Input: Given the observation sample $y$ and basis library $D$.

1. Initialization: Initialize the residual as $r_0 = y, l = 1, \alpha = []$
2. Matching stage: Choose one block best matched to $r_{l-1}$ according to
   \[ i_l = \arg \max_i \| D^T[i] r_{l-1} \|_2, \]
   where $D[i]$ is the $i^{th}$ block of basis library $D$
3. Outlier elimination: Break and return $y$ as an outlier if $l = 1$ and $i_l \neq 1$
4. Estimation stage: Solve the least-squares problem
   \[ \min \| y - \sum_{i \in I} D[i] \alpha_l[i] \|_2, \]
5. Updating stage: Update the residual $r_l = y - \sum_{i \in I} D[i] \alpha_l[i]$.
6. Increment $l$ and go back to step 2, until the $l_2$-norm of residual is below the destination threshold or the maximum number of iterations is reached

Output: The sparse coefficient vector \( \alpha \)
The time complexity of BOMP $\approx \frac{1}{d}$ of OMP if all the whole nonzero entries are clustered by $d$-length blocks identically.

An inherent benefit of the BOMP algorithm comes from its discriminative ability of inferring whether the observed samples are invalid.

We know a priori that a valid observation should be better represented by the target templates rather than the occlusion templates and it should achieve the highest correlation with the target template set.

The matching stage thus can act as a classifier that eliminates the outliers by judging whether the target template set is picked in the first iteration.

It is also possible to shorten the running time of the algorithm if we terminate the loops in an early stage once the observed sample is determined as outliers.
Results
References

Thank you