Analyzing the Subspace Structure of Related Images: Concurrent Segmentation of Image Sets

Lopamudra Mukherjee[1], Vikas Singh[2], Jia Xu[2], Maxwell D. Collins[2]

[1]Mathematics & Computer Science
University of Wisconsin Whitewater
mukherjl@uww.edu

[2]Biostatistics & Med. Info., Computer Sciences
University of Wisconsin – Madison
{vsingh,jiaxu,mcollins}@cs.wisc.edu

Problem: Extract common objects concurrently from a large set of related images oblivious to scale variations.

Motivation

1. Large collection of images of objects are ubiquitous
2. Most current approaches for multi image segmentation are limited to extracting a single similar object across the given image set

   or

   Do not scale well to a large number of images containing multiple objects varying at different scales
3. **Need** an approach with ability to handle multiple images with multiple objects showing arbitrary scale variations

Advantages of the Proposed Approach

- ⋆ No limitations on foregrounds sharing an appearance model or rank constraint on foreground vectors
- ⋆ Permits general non-parametric appearance model compositions of multiple objects at arbitrary scales
- ⋆ Extendable to both unsupervised and supervised settings

The Subspaces of Multiple Object Foregrounds

A new objective for regularizing the coherence among foregrounds of multiple images

Main Ideas

- Create texton histograms for each image where cluster centers with their corresponding covariances define a visual word
- Let $\{m_1, \cdots, m_d\}$ denote histograms for d objects, where for each object l, $m_l \in \mathbb{R}^k$
- Foreground of each image i denoted as a linear combination of object appearances $f^{[i]} = \alpha_1 m_1 + \ldots + \alpha_d m_d$
- Regularize concurrent segmentation of image sets with above subspace constraint

Related Work

- Single Object, two images (*Rother 2006, Mukherjee 2009, Hochbaum 2009*)
- Single Object, Multiple images, Interactive (*Batra 2010*)
- Single Object, Multiple images with scale invariance (*Mukherjee 2011*)
- Others (*Joulin 2010, Kim 2011, Chang 2011, Kim 2012*)

An Unsupervised Model

Foreground appearance vectors for s **images**
$\{F(:,1), \cdots, F(:,s)\} = \{f^{[1]}, \cdots, f^{[s]}\}$ and
Foregrounds sharing common objects expressed as
$F = FC$ where $\text{diag}(C) = 0$

Let $Z^{[i]}$ be the binary matrix constructed from histograms; we get

$$\min_{\mathbf{x},C,\zeta} \sum_i E_{\text{seg}}(\mathbf{x}^{[i]}) + \|\zeta\|^2$$
$$\text{s.t.} \quad \text{diag}(C) = 0, \quad \text{rank}(C) \le \kappa \text{ (a small constant)}.$$
$$F = \hat{F} + \zeta, \quad \hat{F} = \hat{F}C, \quad Z^{[i]}\mathbf{x}^{[i]} = F(:,i),$$

Substituting the low rank requirement with the nuclear norm, we can write an equivalent model as

$$\min_{\mathbf{x},C,\zeta} \sum_i E_{\text{seg}}(\mathbf{x}^{[i]}) + \gamma_1\|F - \hat{F}\|^2 + \gamma_2\|\hat{F} - \hat{F}C\|^2 + \|C\|_*$$
$$\text{s.t.} \quad \text{diag}(C) = 0, \quad Z^{[i]}\mathbf{x}^{[i]} = F(:,i),$$

Algorithm

1. Choose a matrix $\hat{F}$ based on some initialization (e.g., the matrix of all ones).
2. With $\hat{F}$ given, solve
$$\min_{\mathbf{x}} \sum_i E_{\text{seg}}(\mathbf{x}^{[i]}) + \|F - \hat{F}\|^2 \ \text{s.t} \quad \mathbf{x} \in [0,1],$$
   to recover $\mathbf{x}$. Using $\mathbf{x}$, calculate each column of F as $Z^{[i]}\mathbf{x}^{[i]}$.
3. Then, solve the model below to recover $\hat{F}$ and C,
$$\min_{\hat{F},C} \gamma_1\|F - \hat{F}\|^2 + \gamma_2\|\hat{F} - \hat{F}C\|^2 + \|C\|_* \ \text{s.t.} \quad \text{diag}(C) = 0$$
   keeping F fixed.
4. Repeat Steps 2–3 until negligible change in solution.

Properties: Both Step 2 and Step 3 can be solved optimally.

Lemma 1. *The objective value of the relaxed version (above) is non-increasing with each iteration.*

A Supervised Model

1. Previous model needs discriminative backgrounds
2. Instead, use scribble guidance to generate an approximate texton-based appearance model

Two flavors of the problem

A) With precise dictionary

$$\min_{\mathbf{x}^{[i]},\lambda} E_{\text{seg}}(\mathbf{x}^{[i]}) + \gamma\|F(:,i) - \sum_{m_j \in \mathbf{M}} \lambda_j m_j\|^2$$
$$\text{s.t.} \quad F(:,i) = Z^{[i]}\mathbf{x}^{[i]}, \quad \mathbf{x}^{[i]} \in [0,1]$$

Equivalently . . .

$$\min_{\mathbf{x}^{[i]},\lambda} E_{\text{seg}}(\mathbf{x}^{[i]}) + \gamma\|F(:,i) - \text{proj}_{\mathbf{M}}(F(:,i))\|^2$$
$$\text{s.t.} \quad F(:,i) = Z^{[i]}\mathbf{x}^{[i]}, \quad \mathbf{x}^{[i]} \in [0,1]$$

where $\text{proj}_{\mathbf{M}}(F(:,i))$ is the projection of $F(:,i)$ onto the subspace of M, the matrix of object appearances.

Properties: Can be written as Pseudoboolean function in x.

B) With overcomplete dictionary

1. Use a large collection of object appearances, *dictionary* to facilitate the process of segmentation

$$\min_{\mathbf{x}^{[i]},\lambda} E_{\text{seg}}(\mathbf{x}^{[i]}) + \gamma \sum_i \|F(:,i) - \sum_{m_j \in A, A \subseteq D, |A| \le \beta} \lambda_j m_j\|^2$$
$$\text{s.t.} \quad F(:,i) = Z^{[i]}\mathbf{x}^{[i]}, \quad \mathbf{x}^{[i]} \in [0,1]$$

Combinatorial Properties

Let $L(F(:,i), A) = \|F(:,i) - \sum_{m_j \in A} \lambda_j m_j\|^2$, and
$G(F(:,i), D) = L(F(:,i), \phi) - \min_{A \in D, |A| \le \beta} L(F(:,i), A)$,

Observation. Express as $\min E - G$: E is submodular and G is (approx.) submodular. So, $E - G$ is sum of submodular and (approx.) supermodular terms.

Replace supermodular term with approximate modular approximation Ψ: $\Psi(F(:,i), A) = L(F(:,i), \phi) - L(F(:,i), A)$.

Algorithm

1. Solve the function E and get initial estimate for $F_{[t]}$.
2. Solve
$$A_{[t]} = \arg\max_{A \subseteq D} G(F_{[t]}, D).$$
   Since $G(F_{[t]}, D) = \psi(F_{[t]}, A_{[t]})$, we have $E - G(F_{[t]}, D) = E - \psi(F_{[t]}, A_{[t]})$.
3. Solve
$$\min_{\mathbf{x}} E_{seg} - \psi(:, A_{[t]}) \quad \text{keeping } A_{[t]} \text{ fixed.}$$
   Let solution be $\mathbf{x}_{[t+1]}$ and foreground matrix be $F_{[t+1]}$.
4. Repeat Steps 2–3 until negligible change in solution.

Experimental Results

Subspace Cosegmentation of Multiple Objects

Fig. 1: Row 2: Our algorithm. Row 3: Joulin 2010

Cosegmentation with appearance dictionaries

Fig. 2: Results of our algorithm on the iCoseg (cols 1-5) and MSRC (cols 6-8)

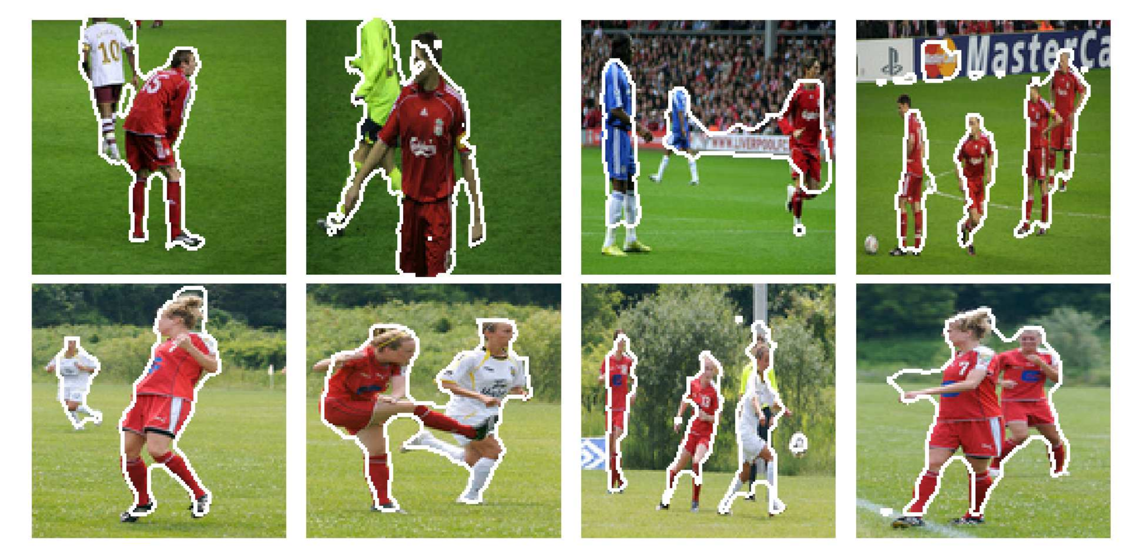

Fig. 3: Results on multi-object Liverpool and Soccer sets

Fig. 4: Row 2: Our algorithm. Row 3: SVM

class	Ours	Vicente 11	Vicente 10	Joulin 2010	class	Ours	Vicente 11	Vicente 10	Joulin 2010
Balloon	**95.17%**	90.10%	89.30%	85.20%	Kite Panda	**93.37%**	90.20%	70.70%	73.20%
Baseball	**95.66%**	90.90%	69.90%	73.0%	Panda	**92.83%**	92.70%	80.00%	84.00%
Brown bear	88.52%	**95.30%**	87.3%	74.0%	Skating	**96.64%**	77.50%	69.9%	82.1%
Elephants	**87.65%**	43.10%	62.3%	70.1%	Statue	**96.64%**	93.80%	89.3%	90.6%
Ferrari	**89.95%**	89.90%	77.7%	85.0%	Stonehenge1	**92.67%**	63.30%	61.1%	56.6%
Gymnastics	**92.18%**	91.70%	83.4%	90.9%	Stonehenge2	84.87%	**88.80%**	66.9%	86.0%
Kite	**94.63%**	90.3%	87.0%	87.0%	Taj Mahal	**94.07%**	91.1%	79.6%	73.7%

Table 1: Segmentation accuracy for iCoseg dataset.

European Conference on Computer Vision (ECCV), 2012.