Learning Similarity Metrics for Dynamic Scene Segmentation

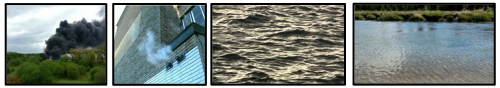
Damien Teney, Matthew Brown, Dimitry Kit, Peter Hall. University of Bath, UK

Motivation

Goal: improve video segmentation using motion cues

Challenge: dynamic textures, non-rigid objects/motions (smoke, water, foliage, fire, ...)

Usual approach: ~~optical flow~~ + parametric ~~motion models~~



Proposed approach:

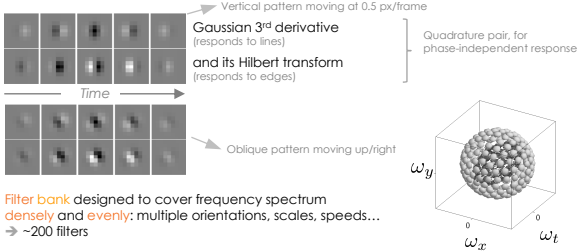
- **Spatio-temporal filters to identify motion and changes of appearance**
  - Capture rigid/non-rigid motions, brightness changes, flickering effects, ...
  - Local measurements: no commitment to early motion interpretation
  - Region descriptors from histograms of filter responses
  - Used as features in model-free, unsupervised segmentation
- **Add supervision to graph-based segmentation via distance learning**
  - Large-margin metric learning: separates segments to merge / keep apart
  - Constraints from ground truth segmentations / semantic examples

Filter-based motion features

Like 2D filters identify oriented structures (edges) in 2D images

3D filters are applied on the video volume of stacked frames

Steered in 3D to particular orientations / velocities



Filter bank designed to cover frequency spectrum densely and evenly: multiple orientations, scales, speeds...

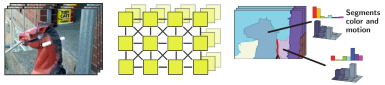
→ ~200 filters

Segmentation framework

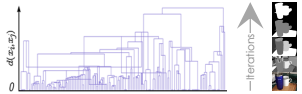
Graph-based segmentation, **regions described by color + motion histograms**

Grundmann, Kwatra, Han, Essa, Efficient hierarchical graph-based video segmentation, CVPR, 2010.

Graph of adjacent segments
Initially 1 pixel = 1 node
Edge weight = dissimilarity (distance) between nodes



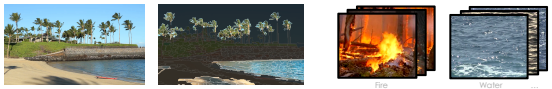
Agglomerative clustering
→ Hierarchical tree of segments



Supervised segmentation via distance learning

Unsupervised segmentation = grouping regions with similar features

Idea: learn this similarity measure, from 2 types of examples:

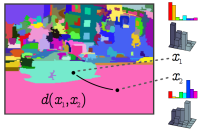


Hand-drawn segmentations

Semantically-labelled examples

Replace the unsupervised histogram distance with a learned, generalized Mahalanobis distance:

$$d_L^2(x_1,x_2) = (x_1-x_2)^\top L^\top L\,(x_1-x_2)$$

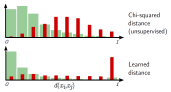


Constraints

Training segments: features x_i
Pairwise annotations $y_{ij} = +1/-1$
$d_L(x_i,x_j) \ll d_L(x_k,x_l)$
$\forall\, i,j,k,l$ s.t. $y_{ij}=+1,\ y_{kl}=-1$.

Large-margin objective

$$\arg\min_{L,t} \textstyle\sum_{i,j} \max\left[1-y_{ij}\left(t-d_L^2(x_i,x_j)\right),\ 0\right]$$



Also integrate dimensionality reduction via linear projection

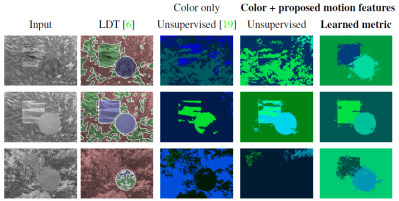
$$d_L^2(x_1,x_2) = (x_1-x_2)^\top L^\top L\,(x_1-x_2) = \|\,(Lx_1-Lx_2)\,\|_2^2$$

Non-square matrix $L \in \mathbb{R}^{p\times d},\ p<d$

Projects $x_i \in \mathbb{R}^d$ to a space of lower dimension $\mathbb{R}^p$

Experiments

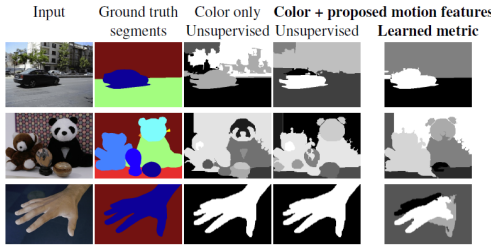
Dynamic texture segmentation (SynthDB benchmark)



Method		Avg. Rand (%)
Features	*Metric*	
Color	Unsupervised	59.9
Color + motion	Unsupervised	72.7
Color + motion	Learned, logistic regression	77.1
Color + motion	Learned, ITML [10]	86.4
Color + motion	Learned from manual segm.	89.7
Color + motion	Learned from manual segm. and semantic labels (*pwf*)	**90.2**
GPCA [24]		55.4
LDT [6] (with manual initialization)		89.4
DTM (CS) [4] (static segments)		82.5
(LBP/WLD)$_{TOP}$ [7]		88.4

Complex natural scenes (Dyntex dataset)

It still works with rigid motions, too ! → More generally applicable than e.g. optical flow



Work supported by EPSRC Grant EP/K02339X/1.